READER-TEST INTERACTIONS: AN EXPLANATORY ITEM RESPONSE STUDY ON READING COMPREHENSION

By

Ping Wang

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Literacy Studies

Middle Tennessee State University

May 2021

Dissertation Committee:

Dr. Jwa K. Kim, Chair

Dr. Amy M. Elleman

Dr. Eric L. Oslund

ACKNOWLEDGEMENTS

I am extremely grateful to my advisor, who is also the chair of my dissertation committee, Dr. Jwa Kim. During my four years of study, he has provided me with great academic guidance and spiritual encouragement. His statistical knowledge has helped me lay a solid foundation for my future research career, and his virtuous character will have a profound impact on my future life.

I would like to express my sincere thanks to Dr. Amy Elleman and Dr. Eric Oslund, and all my other professors for their thoughtful guidance and continued support throughout this process. Their erudition and outstanding teaching skills broadened my mind and gave me the preparation I needed to complete this dissertation. I really appreciate the time and assurance they gave to me.

I would give my special thanks to Dr. Casey Brasher for sharing her data. Without her data, I would not have been able to complete this research project. I want to thank Dr. David Francis, Dr. Paulina Kulesz, and Dr. Sun Lili for their help in the data analysis process, and my classmates, Jennifer Francois and Nicole Crouch for their corrections and comments on my dissertation. I would also like to send my thanks to my friends, Bingshi Zhang, Xiao Tan, Yu Qiao, and other classmates for giving me so much care, comfort, company, and joy in my Ph.D. life.

Lastly, I would like to express my deep appreciation to my parents and my sister. Their love, support, and encouragement are always the sources of my courage and strength, making me chase my dream with no worries or concerns. After five years apart, it is time to reunite.

ABSTRACT

According to the RAND model framework, reading comprehension test performance is influenced by readers' reading skills or reader characteristics, test properties, and their interactions. However, little empirical research has systematically compared the impacts of reader characteristics, test properties, and reader-test interactions across different standardized tests. The present study used the explanatory item response approach to investigate the reader-test interactions in two commonly used standardized tests: *Gates-MacGinitie Reading Test-4th edition* (GMRT-4; MacGinitie et al., 2000) and *Wechsler Individual Achievement Test- 3rd Edition Reading Comprehension* (WIAT-III; Wechsler, 2009). Five reader characteristics scores (i.e., decoding, pseudoword reading, vocabulary, fluency, morpho-syntactic knowledge) of 89 fourth graders were obtained. Six test properties (i.e., mean sentence length, mean log word frequency, referential cohesion, deep cohesion, genre, and question type) of the two tests were measured by The Lexile Framework for Reading (Schnick & Knickelbine, 2007) and Coh-Metrix Text Common Core Ease and Readability Assessor (Coh-Metrix-TERA; Graesser et al., 201). Genres were coded as narrative texts and expository texts, and question types were classified into literal questions and inferential questions. Explanatory item response models (EIRM) treated both reader characteristics and test properties as random variables. Results indicated in both GMRT-4 and WIAT-III, fluency and vocabulary were the most crucial reader characteristics over other reading skills. For GMRT-4, lower mean sentence length, higher referential cohesion, and higher deep cohesion made the passage easier to understand, and expository texts were more difficult than narrative texts. For WIAT-III, inferential questions were more challenging than literal questions. Three significant reader-test interactions were found in GMRT-4 between vocabulary and referential cohesion, vocabulary and word frequency, and referential cohesion and question type. Higher-level vocabulary students had better comprehension performance in passages with low referential cohesion and high word frequency in GMRT-4. Literal questions were easier than inferential questions for low referential cohesion texts, but the result was reversed for high referential cohesion texts. Through this study, fourth graders' reading performance in the two standardized tests was better understood, and limitations and implications for practice and future research were discussed.

Keywords: reading comprehension, test properties, reader characteristics

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Reading comprehension is a nonunitary construct but involves complicated cognitive skills of readers and text features of reading materials. The scope of comprehension has been addressed by numerous reading theories and models, among which, bottom-up processing models such as the Construction-Integration model (C-I; Kintsch & van Dijk, 1978), Simple View of Reading (SVR; Gough & Tunmer, 1986), and Direct and Inferential Mediational model (DIME; Cromley & Azevedo, 2007) accentuate the importance of component skills during reading, while top-down models such as the Structural-Building model (Graesser et al., 1997) emphasize the active role of prior knowledge of readers. Although various reading theories and models have been proposed, they were not apparent in the most commonly used and referenced standardized reading comprehension assessments (Kintsch & Kinstch, 2005).

Reading comprehension tests are constructed to detect students' reading difficulties, monitor students' reading progress, and help develop or test reading theories (Cain & Oakill, 2006). However, many reading tests measure different reading skills and are not always interchangeable with each other in evaluating students' reading ability (e.g., Betjemann et al., 2011; Nation & Snowling, 1997). The inconsistency among reading tests brings challenges in classifying students at different levels (Keenan & Meenan, 2014; Rimrodt et al., 2005), thereby may affect appropriate intervention decisions and classroom instruction.

Student's test performance is impacted by both external and internal factors such as reading test construction and individual differences. With respect to test development, people could have different concepts toward reading test construction before fully understanding the reading process. For instance, researchers influenced by behaviorism focused more on the psychometric criteria in reading tests, while those influenced by constructivism addressed the cognitive processes being assessed by test questions (Pearson, 2000; Waston, 1913). Comprehension tests also differ widely in test formats such as cloze tests, multiple-choice questions, passage lengths, or question types so that they could tap different component skills (Francis et al., 2005; Keenan et al., 2008; Nation & Snowling, 1997). In terms of individual differences, reader characteristics such as word reading ability and vocabulary are significant predictors of reading test performance (Cain et al., 2004; Keenan et al., 2008; Keenan & Meenan, 2014). These reader characteristics could also interact with test properties such as passage features and question types, which would produce moderating effects on comprehension outcomes (Best et al., 2008; McNamara et al., 2011).

Although the impacts of reader characteristics and test properties on test performance have been noted, they are mostly investigated as two isolated lines. For instance, researchers examined the variance explained by reader characteristics in different tests (e.g., Cain et al., 2004; Cutting & Scarborough, 2006; Keenan et al., 2008) or exclusively explored how test properties (text complexity and task features) influence the passage difficulty (e.g., Bormuth, 1969; Klare, 1984). Despite some studies examining reader-test interactions, the adoption of total raw scores based on classical test theory (CTT; Lord & Novick, 1968) could average over the influence of different reading skills and text features on reading outcomes (Best et al., 2008; McNamara et al., 2011). The extant research falls short of examining how reader characteristics, test properties, and their interactions simultaneously influence the performance on standardized tests at the item level. Also, no research systematically compares these influences across different standardized tests. Therefore, this study focuses on explaining reading comprehension performances in different standardized tests through item-level scores.

Theoretical Framework for Reading Comprehension

Researchers have emphasized it is crucial to construct and interpret reading assessments within the framework of reading theories or models (Duke, 2005; Kintsch and Kintsch, 2005; Pearson & Hamm, 2005). Components-based models such as the SVR (Gough & Tunmer, 1986) and the DIME model (Cromley & Azevedo, 2007) emphasize the role played by component skills such as decoding, vocabulary, and fluency during reading. Information processing models like the landscape model (van den Broek et al., 2005; van den Broek et al., 1999) and the structural-building model (Gernsbacher, 1997; Gernsbacher et al., 1990) try to delineate the process of mental representation of a text. Among various reading models, the RAND model is regarded as a comprehensive model, which covers almost all reading-related components. The current study investigated reading comprehension performance under the framework of the RAND model.

The RAND model defines reading comprehension as a "process of simultaneously extracting and constructing meaning through interaction and involvement with written language" (Snow, 2002, p. 11). The RAND definition contains three elements in the reading process: the reader, the text, and the activity or purpose for reading. The three elements interrelate with each other, and such interrelationship takes place under a larger sociocultural context.

Reader characteristics include an array of capacities and abilities related to comprehending a text, such as cognitive skills, motivation, and knowledge (Snow, 2002).

From the perspective of readers, basic reading skills such as decoding, vocabulary, fluency, and morpho-syntactic knowledge all make significant contributions to successful reading comprehension (Cain et al., 2001; Carlisle & Stone, 2003; Cutting & Scarborough, 2006; Elleman et al., 2009; Elleman et al., 2017; Keenan et al., 2008; Oslund et al., 2018; Yovanoff et al., 2005). Reader characteristics also vary across individuals and function differently over time in texts exposed to readers (Catts, 2018; Snow, 2002).

As the information carrier, text features play an essential role in the reading process. The variability of text features is usually examined from different categories and dimensions. Sentence difficulty such as vocabulary and syntax, discourse genres such as narrative or expository text, and discourse structure such as rhetoric and coherence are examples of text features (Snow, 2002; White, 2010). Each of the variables could impose difficulties on comprehension so that readers might find it easier to comprehend a text with short passage length and a familiar topic but feel challenged to understand materials with difficult vocabulary and syntax (McNamara et al., 2011). It has been acknowledged that successful reading comprehension depends on the match between text features and reader characteristics.

The third element is reading activity which refers to the behavior that readers interact with the text with a particular purpose (Snow, 2002). The purposes or aims of reading may change for pre-reading, during-reading, and after-reading processes. Such activities often occur under specific contexts, in which contextual factors such as cultural group or discourse community, the school building, and classroom instruction are usually considered (Geske & Ozola, 2009).

Within the RAND model framework, readers in the test scenario are test takers who bring their reading skills such as decoding, vocabulary knowledge, and fluency to understand passages and answer test questions. The text features are termed as passage features including word frequency, sentence length, cohesion, and genre. Reading activity in this dissertation is regarded as answering different question types such as literal questions and inferential questions. Passage features and question type were together termed as test properties throughout this dissertation. In this way, the explanation of the reading comprehension process and reading performance in the current study becomes the exploration of the interrelationship among reader characteristics, passage features, and question types.

Reader Characteristics Assessed in Reading Comprehension Tests

As mentioned before, reader characteristics under the RAND model include an array of abilities or skills related to comprehending a text. These different component reading skills have been examined in various tests. Although there are inconsistencies, the moderate correlations among various reading tests indicate some essential skills they are typically assessing (Keenan & Meenan, 2014; Kendeou et al., 2012; Kendeou et al., 2014). In this study, four reading skills were investigated: decoding, vocabulary, fluency, and morpho-syntactic knowledge.

As for decoding, Cutting and Scarborough (2006) compared three commonly used reading comprehension tests (*Gates- MacGinitie Reading Tests-Revised* [GMRT]; MacGinitie et al., 2000; *Gray Oral Reading Test- Third Edition* [GORT-3], Wiederholt & Byant, 1992; *Wechsler Individual Achievement Test* [WIAT], Wechsler, 1992) and found decoding skill accounted for unique variance in all three tests. Nevertheless, WIAT was more influenced by students' decoding skills with higher variance than GMRT and GORT-3. Similarly, Keenan et al. (2008) also demonstrated that decoding accounted for most of the variance in four tests: GORT-3, *Qualitative Reading Inventory*–3 (QRI-3) (Leslie & Caldwell, 2001), *Woodcock-Johnson Passage Comprehension subtest* (WJPC) from the *Woodcock-Johnson Tests of Achievement–III* (Woodcock et al., 2001) and *Peabody Individual Achievement Test* (PIAT) (Dunn & Markwardt, 1970). A factor analysis showed that the decoding factor loaded more on PIAT and WJPC. The role of reading skills in reading performance needs more evidence with the inclusion of various standardized tests.

In addition to decoding, vocabulary was also widely investigated in research and was consistently examined as a robust and constant predictor of reading comprehension irrespective of grade level (e.g., Oslund et al., 2016; Verhoeven et al., 2011; Yovanoff et al., 2005). Not only vocabulary quality (i.e., knowledge about the word's form, meaning, and use) but also the size of vocabulary amount has a strong association with reading comprehension (Ouellette, 2006; Perfetti & Hart, 2001; Verhoeven et al., 2008). Research also showed that vocabulary influenced inference-making, word decoding, and fluency (Kim, 2017; Ouellette, 2006; Yovanoff et al., 2005). Nevertheless, a meta-analysis found that vocabulary instruction was more effective in comprehending texts with custom measures but less effective with standardized measures (Elleman et al., 2009). It is necessary to examine whether vocabulary skills have interactions with test properties of standardized reading tests.

Additionally, oral reading fluency is frequently explored by researchers, especially for elementary school readers. In a longitudinal study, reading fluency was found to be a significant predictor for reading comprehension, but its effect diminished over years or grades (Yovanoff et al., 2005). According to Kim et al. (2010), through grade 1 to grade 3, oral reading fluency had the strongest prediction power in reading comprehension in Grade 3. Although the effect of oral reading fluency on comprehension declines afterward, it is still necessary to take fluency level into account in the elementary stage when explaining the reading achievement.

At last, morpho-syntactic knowledge is considered useful in constructing the meaning of texts. According to Cutting and Scarborough (2006), syntactic knowledge contributed uniquely to reading comprehension with a variance of 1-5% for 7-15-year-old students. In longitudinal studies, syntactic knowledge contributed 4-24% of the variance of reading comprehension for elementary students (e.g., Demont & Gombert, 1996; Muter et al., 2004).

Test Properties in Reading Comprehension Assessment

Previous research suggested that test properties could be predictors for reading comprehension performance. Test properties are typically examined from the perspectives of passage features (e.g., sentence length, word frequency, genre, cohesion) and question types (e.g., literal questions, inferential questions). Passages with shorter sentence lengths and higher word frequency are easier to understand (Kincaid et al., 1975). Passage genres, such as narrative texts and expository texts, have different demands for cognitive processing and could exert a notable effect on comprehension. Previous research demonstrated that narrative texts were less challenging than expository texts (Best et al., 2008; McNamara et al., 2011). Another commonly examined passage feature is cohesion which refers to how the events and concepts are connected within a text (McNamara et al., 2012). High cohesion texts have been shown to be easier to comprehend than low cohesion texts, indicating cohesion influences a text’s readability (MaNamara et al., 2011). Among the various sources of cohesion, referential cohesion and deep cohesion were two important indices for text complexity (Graesser et al., 2011).

Apart from passage features, question type is another test property of reading tests to be considered. Literal and inferential questions are common question types that are widely used in reading assessments. Literal questions require students to recall explicitly the information presented by the text. In contrast, inferential questions assess students’ ability to draw conclusions by integrating pieces of text information or combining text information with readers’ prior knowledge (Eason et al., 2012; Miller et al., 2014). It has been found that inferential questions are more difficult than literal questions (Basaraba et al., 2012; Muijselaar et al., 2017), but question types could interact with passage genres, which might influence the difficulty of questions (Eason et al., 2012).

Since there are various test properties, it is reasonable to speculate interactions between these variables. In a study examining the interactions among test properties, Eason et al. (2012) found an interaction between genre and question types. For narrative text, initial understanding questions (literal questions) were significantly easier than interpretation questions. There was an opposite effect for expository text: initial understanding questions were significantly more difficult than interpretation questions. Therefore, test performance interpretation should not only explore the influence of test properties separately but consider the possible interactions as well.

Explanatory Item Response Models (EIRM)

The variability of reader characteristics and test properties demonstrates the complexity in interpreting reading test performance. The traditional way of viewing test scores is based on classical test theory (CTT, Lord & Novick, 1968), which concerns the sum score or total score of all test items. One weakness of total score is that it will aggregate the influence of person abilities and item features. In contrast, item response theory (IRT) focuses on the item level information and assumes the correct response of an item is a function of both person parameters and item parameters (Hambleton & Van der Linden, 1982; Thomas, 2011). Explanatory item response models (EIRM) expand the application scope of IRT models by allowing the addition of predictive variables pertaining to both persons and items. In this way, EIRM obtained its explanatory nature by describing item performance with regard to other variables (De Boeck & Wilson, 2004). Typically, EIRM deals with data categorical in nature, such as dichotomous data and polytomous data (Kim & Wilson, 2020; Kubinger, 2009).

The most significant advantage of applying EIRM lies in its direct modeling and estimation of predictive variables, or “one-step method,” though in practical applications, it is also possible to use the “two-step” method. For instance, the first step is to use the IRT model to obtain the person and item parameter estimates, such as reader ability scores and item difficulty parameters. The second step is to regress the parameter estimates on explanatory variables, such as reader characteristics and test properties (Briggs, 2008; Hartig et al., 2011). EIRM, however, is superior to the two-step method in reducing measurement error and providing convenient estimation for more complex design (Debeer & Janssen, 2013).

In the reading comprehension test instance, EIRM makes it possible to simultaneously examine how reader characteristics and test properties jointly explain students' test performances (Kulesz et al., 2016). It seeks to explain reading test performance by a function of reader characteristics which denotes the reader's position on the latent trait continuum, and a function of test properties which refers to the item's position on the difficulty continuum. The explanatory IRT approach can incorporate external design variables concerning both the reader ability and item difficulty into measurement models. It also allows for random variations in the two dimensions, making the results generalizable to persons and tests.

Purpose of the Study and Research Questions

In recent years, the explanatory IRT approach has been applied in the reading area by several researchers, and its flexibility helps detect different interactions between reader characteristics and test properties (Kulesz et al., 2016; Miller et al., 2014; Spencer et al., 2018). Nevertheless, for the most part, reading measures used in previous studies were usually designed by researchers for specific study purposes, and the exploration for commonly used standardized tests has not been sufficient. The study by Kulesz et al. (2016) used EIRM to investigate the reader-text interaction in GMRT among middle school and high school students separately. However, it is still unknown what interactions will happen in GMRT for elementary students. Moreover, since the two groups in their study were tested with different items, comparisons between the two groups were hard to make. Given the inconsistency across various tests as mentioned before, it is necessary to investigate whether the reader-test interactions in one standardized test are comparable to others using within-group designs.

To fill the current gap, this study examined the effect of reader characteristics, test properties, and reader-test interaction on the test performance of fourth grade students. Fourth grade is a critical period in the elementary stage because students are transiting from the phase of “learning to read” to “reading to learn,” and reading passages are incorporating more forms such as essays and expository texts (Toste & Ciullo, 2016; Wanzek et al., 2010). Studying the reading test performance of fourth graders will help understand the reading process and inform better test results’ interpretation and classroom instruction. In the current study, two standardized tests were included: *Gates-MacGinitie Reading Tests- 4th Edition* (GMRT-4; MacGinitie et al., 2000) and *Wechsler Individual Achievement Test- 3rd Edition Reading Comprehension* (WIAT-III; Wechsler, 2009).

EIRM was utilized to examine how reader characteristics, test properties, and the reader-test interactions influence the reading performance in the two tests. Four classes of models were estimated for GMRT-4 and WIAT-III test data. Model 0 was an unconditional model without any predictors to estimate the residual variance on reader ability and item difficulty of GMRT-4 and WIAT-III; Model 1 added reader characteristics (i.e., decoding, pseudoword decoding, vocabulary, fluency, and morpho-syntactic knowledge) into Model 0 to estimate which reader characteristics were important for the reader ability; Model 2 added test properties (passage features and question type variables) into Model 0 to estimate which test property best predicted the item difficulty; Model 3 added both reader characteristics and test properties to examine potential reader-test interaction effects on test performance. The following research questions for the two sets of models of GMRT-4 and WIAT-III were addressed:

1) How are reader characteristics (i.e., word reading and decoding, vocabulary, fluency, morpho-syntactic knowledge) of fourth grade students related to the test performances in GMRT-4 and WIAT-III?
2) How are test properties (i.e., mean sentence length, mean word frequency, referential cohesion, deep cohesion, genre, question types) related to the test performances in GMRT-4 and WIAT-III?
3) Is there any moderating effect of reader characteristics and test properties on the test performances in GMRT-4 and WIAT-III? Based on previous research, four interactions were investigated between: vocabulary and referential cohesion, vocabulary and word frequency, referential cohesion and question type, and genre and question type.

CHAPTER II: LITERATURE REVIEW

This chapter reviewed relevant literature on the effects of reader characteristics, test properties, and reader-test interactions in reading comprehension performance. Reader characteristics were mainly concerned about the elementary stage, especially in fourth graders; test properties included both passage features and question types in reading tests; interaction effects were not limited to reader-test interactions but included passage-question interactions in reading tests. In addition, literature about the theoretical framework of Rasch EIRM, the manipulation of binary data, and its advantages of application in reading comprehension were reviewed.

Impact of Reader Characteristics on Reading Ability

Reader characteristics under the RAND model can be regarded as an umbrella term covering a wide range of abilities and skills (Snow, 2002). Reader characteristics explained a large amount of variance in reading performance and were important sources of variation among good readers and poor readers (Cain et al., 2001; Cain et al., 2004; Cutting & Scarborough, 2006; Keenan et al., 2008). The present study mainly focused on the influence of decoding, vocabulary, fluency, and morpho-syntactic knowledge on reading comprehension in the elementary stage.

Word Reading and Decoding

In most component-based models, decoding was placed in a significant position. According to SVR, reading comprehension was the product of decoding (or word reading) and language comprehension (Gough & Tunmer, 1986; Hoover & Gough, 1990). The automaticity of word reading could directly influence how many cognitive resources were allocated to text meaning construction (e.g., Jenkins et al., 2003; Perfetti, 1985; Walczyk, 2000).

In some reading tests, decoding explained more variance than other reading skills. For instance, Cutting and Scarborough (2006) compared three popular reading comprehension tests: WIAT, GMRT, and GORT-3. They had children in Grades 1 to 10 take the three tests and measured students' reading skills and cognitive characteristics such as decoding, oral language-lexical skill, oral language-sentence processing, reading speed, rapid serial naming, verbal memory, full-scale IQ, and attention. The results showed decoding skills and oral language accounted for significant variance across all three tests. Nevertheless, in WIAT, decoding accounted for higher unique variance (11.9%) than oral language (9%), and this amount of variance of decoding was also higher than that in GMRT (6.1%) and GORT-3 (7.5%). Through factor analysis, researchers found that PIAT and WJPC were more decoding-related, while GORT-3 and QRI-3 were more listening comprehension-related (Betjemann et al., 2011; Keenan et al., 2008). Similarly, Nation and Snowling (1997) studied two commonly used reading comprehension tests in Britain -- Neale Analysis of Reading Ability (NARA; Neale, 1989) and Suffolk test (Hagley, 1987). They discovered Suffolk was more related to the single word reading ability, while NARA was more dependent on listening comprehension.

However, longitudinal studies on the relationship between decoding and reading comprehension presented a mixed picture. Some researchers argued that the influence of decoding attenuated from the beginning to the end of evaluation points (e.g., Abbott et al., 2010; Burgoyne et al., 2011; Kim et al., 2011; Kim et al., 2012) while others found such relationship kept relatively stable between the two evaluation time points (e.g., Cain et al., 2004; Cain & Oakhill, 2011; Compton et al., 2008; Oakhill et al., 2003). A meta-analysis by García and Cain (2014) concluded a strong correlation between decoding and reading comprehension across all ages and the variability in different studies attributed to some moderators. These moderators included two reader characteristics (i.e., students' age and listening comprehension level) and three test characteristics (i.e., text genre, the help provided for decoding, and loud reading). Remarkably, the moderator effect of test characteristics was more conspicuous for younger readers. Nevertheless, McNamara et al. (2011) did not find any interaction effect between decoding and genre, which indicated the effect of genre and cohesion did not rely on decoding skills. In short, the impact of decoding on reading comprehension performance for fourth graders who are in the transitional reading stage is still worthwhile to investigate across different tests.

Vocabulary

As Davis (1944) stated, "It is clear that word knowledge plays a very important part in reading comprehension" (p. 191). Various sources of studies, such as descriptive analysis, correlational studies, and empirical studies of vocabulary instruction, have demonstrated that vocabulary was a significant factor in reading comprehension (e.g., Baumann, 2009; Bloom, 1976; Cromley & Azevedo, 2007; Davis, 1944; Elleman et al., 2009; Ouellette, 2006; Ouellette & Beers, 2009; Thorndike, 1917). Vocabulary explained unique variance in reading comprehension beyond decoding and listening comprehension (Braze et al., 2007).

Previous research examined the relationship between vocabulary and reading comprehension from two aspects: vocabulary breadth (e.g., how many words are known;

Tannenbaum et al., 2006) and depth (e.g., how well the meanings are known; Ouellette, 2006). According to the Lexical Quality Hypothesis (LQH), high quality lexical representation facilitated reading skill and comprehension (Perfetti & Adolof, 2012; Perfetti & Hart, 2001; Perffeti & Stafura, 2014). Empirical studies showed that skilled readers were different from less skilled readers on both breadth and depth of word knowledge (Braze et al., 2007; Cain & Oakhill, 2014; Oakhill & Cain, 2012; Ouellette, 2006; Tannenbaum et al., 2006).

Also, the contribution of vocabulary to reading comprehension has been shown to increase across grades. Ouellette and Beers (2009) investigated the effect of oral vocabulary (indicated by vocabulary breadth and depth) in Grade 1 and Grade 6, respectively. Multiple regression model revealed for Grade 1, vocabulary breadth and depth did not contribute significantly to the variance of reading comprehension. In Grade 6, although vocabulary depth was still not a significant predictor, vocabulary breath explained a unique variance of 15.3%. The results demonstrated the increasing contribution of vocabulary in reading comprehension across grades. Similarly, Kim (2020) examined the relationship between various component skills and reading comprehension through longitudinal data sets from Grade 2 to Grade 4. The findings showed that the indirect effect of vocabulary on reading comprehension increased from .21 in Grade 2 to .38 in Grade 4, which indicated that vocabulary became gradually influential in Grade 4. The researchers also pointed out that it was likely because of the addition of expository texts requiring a higher vocabulary demand in Grade 4. Therefore, it is of interest to see how vocabulary influences the reading test performance of fourth graders and potential interactions between vocabulary and test properties.

Fluency

Reading fluency was defined as “the ability to read text quickly, accurately, and with proper expression” (National Reading Panel, 2000, p. 5)”, and it consists of three key elements: accuracy, rate, and prosody (Hudson et al., 2005). Large amounts of studies have explored the relationship between fluency and reading comprehension by direct modeling of fluency and comprehension, exploring its influence on other reader skills, and investigating the intervention effect. It is generally accepted that fluency contributed uniquely to reading comprehension (e.g., Fuchs et al., 2001; Kim et al., 2010; Kim et al., 2011; Price et al., 2015; Veenendaal et al., 2015; Zimmermann et al., 2019).

Despite a diminished effect over grades (Yovanoff et al., 2005), for fourth graders, oral reading fluency has been shown to continually be positively related to reading comprehension performance (Danne et al., 2005; Jenkins et al., 2003). Sabatini et al. (2018) did a secondary analysis of the relation between oral reading fluency and reading comprehension by a large data set of 140,000 fourth graders collected for the 2002 NEAP study of oral reading (Danne et al., 2005). In this study, oral reading was indicated by three performance scores: rate (words per minute, WPM), accuracy, and prosody. The results demonstrated that reading rate (WPM) was the strongest predictor of reading comprehension, followed by accuracy and prosody. Nevertheless, considering the complex process of reading comprehension, the authors suggested not teaching speed reading alone to improve reading comprehension but adopted evidenced-based fluency approaches. Sabatini et al. (2018) clearly demonstrated oral reading fluency explained variation of reading comprehension at a fourth-grade level, and reading rate and accuracy could be used together (i.e., words read correctly per minute) as indicators for reading fluency.

Given the inconsistency of various reading tests, reading fluency might account for different variances for reading comprehension across tests. Kang and Shin (2019) investigated how decoding and reading fluency predicted the reading comprehension achievement for fourth graders in three reading comprehension tests: GMRT (MacGinitie et al., 2000), Test of Silent Reading Efficiency and Comprehension (TOSREC; Wagner et al., 2010), and WJPC (Woodcock et al., 2001). The results showed that decoding and reading fluency accounted 8.1% of the variance for GMRT, 22.5% of the variance for TOSREC, and 43.3% of the variance for WJPC. Reading fluency contributed uniquely 3.9% of variance to GMRT, 4.5% to the TOSREC, and 1.9% to WJPC.

Compared with decoding, fluency presented relatively similar influences across different tests. As the author noted, it could be due to the less sensitivity of reading fluency to passage format and response format in the upper elementary grades. Other researchers pointed out that reading fluency was negatively related to text difficulties, which means with the increase of text difficulty, reading fluency, such as accuracy, rate, and prosody decreased (Amendum et al., 2017; Cheatham et al., 2014; Young & Bowers, 1995). Therefore, further examination of the impact of fluency with the inclusion of passage features would be helpful for understanding whether fluency has stable influences on different standardized tests.

Morpho-Syntactic Knowledge

Morphological knowledge refers to the ability to identify and manipulate the smallest phonological units – morphemes – within words, in other words, the word formation rules (Carlisle, 1995; Nagy et al., 2014). Syntactic knowledge refers to the ability to “reflect on and manipulate the order of words in a sentence” (Nagy & Scott, 2000, p. 275). Morphological and syntactic awareness were often accessed at the same time by derivational tasks (Carlisle, 2000; Kuo & Anderson, 2006; Nagy et al., 2006; Tong et al., 2014). For instance, the student might be presented a word “teach” and then be required to complete the sentence: She is a good___. To answer this question, the student first should have the syntactic knowledge that the blank needed a noun; and then he or she should have morphological knowledge that a nominal suffix -er was needed to denote a person. A meta-analysis demonstrated that morphological interventions were more effective in the elementary period than middle and high school (Goodwin & Ahn, 2013).

Previous research has consistently agreed that morpho-syntactic knowledge significantly predicted reading comprehension among elementary students (e.g., Bryant et al., 2000; Carlisle, 2003; Carlisle & Stone, 2003; Tong et al., 2014). Berninger et al. (2010) conducted growth curve analysis to investigate the growth of phonological, orthographic, and morphological awareness in the elementary stage by two longitudinal data sets: first to fourth grade and third to sixth grade. The results showed that in the first three grades, morphological awareness gained its steepest growth but did not reach the growth ceiling. In grades four and above, there was some additional growth, which indicated that the morpho-syntactic knowledge had a longer developmental span compared to phonological awareness.

Additionally, for fourth graders, morpho-syntactic knowledge could make significant contributions to reading comprehension outcomes. Using structural equation modeling in fourth and fifth grade, Nagy et al. (2006) found the correlation between morphological awareness and reading comprehension was .76, much higher than phonological short-term memory (*r* = .27) and phonological decoding (*r* = .38). Despite a high correlation (*r* = .85) between morphological awareness and vocabulary, morphological awareness was a significant predictor for reading comprehension outcome in fourth and fifth grades after controlling for vocabulary. The authors pointed out the strong predictive power for morphological awareness in this period might be due to morphological awareness in syntactic parsing. Therefore, using morpho-syntactic knowledge together as an indicator would provide a clearer picture of the nature of the relationship between the morpho-syntactic domain and reading comprehension.

Impact of Test Properties on Test Difficulty

In standardized reading comprehension tests, students' reading ability is typically assessed by the combination of passage reading and question answering. Test properties such as passage features and question types can influence the test difficulty or the probability of a correct question answering (Amendum et al., 2017; Benjamin, 2011). The current research is mainly concerned with selected passage features (i.e., sentence length, word frequency, cohesion, and genre) and question types (i.e., literal and inferential questions) with a discussion of their impacts on test difficulty for elementary students.

Passage Features

Passage features in the current study have the same meaning as the concept of text complexity in the study of Mesmer et al. (2012), which refer to factors that independently influence the readability of a text regardless of reader characteristics and question types. Since a large number of high school graduates were not meeting the reading requirement of colleges and career, Common Core State Standards (CCSS; National Governors Association Center for Best Practices, 2010) and Council of Chief State School Officers suggested increasing the text complexity for reading instruction of Grades 2-12. Researchers have paid much attention to factors that influence text complexity or text readability (e.g., Amendum et al., 2017; Benjamin, 2011; Frantz et al., 2015). The following section reviewed these factors, such as sentence length, word frequency, cohesion, and genre.

Sentence Length and Word Frequency. Traditionally, researchers used the average sentence length and the average number of syllables or word length to predict text readability (Kincaid et al., 1975). Longer sentence length and word length characterized a more difficult text. It should be noted that word length was often utilized as an indicator for word frequency (McNamara et al., 2012).

Adopted by CCSS for grade level reading, Lexile Framework for Reading is a popular automated analysis tool for digitized texts. According to Stenner et al. (2006), the readability of the text in Lexile was predicted by the mean sentence length (MSL) and the mean log word frequency (MLWF). Sentence length and word frequency were proxies for semantic and syntactic components on text complexity. In Lexile, MLWF refers to "the logarithm of the number of times of a word appears in each 5 million words of a corpus of nearly 600 million words" (Lennon & Burdick, 2004). It should be noted that word frequency is not the number of occurrences of a word in a particular text or paragraph, but the frequency that a word occurs in a corpus of 600 million words which is used by the Lexile. However, critics pointed out the indices of Lexile were typical surface-level predictors or unidimensional readability metrics which may not capture the full picture of the reading process, and textural level indicators such as textual cohesion should be considered (Adams, 2001; Graesser et al., 2014; Kamil, 2001; Larson, 2001; McNamara et al., 2002). Thus, in the current study, apart from word level and sentence level predictors, textual level indicators such as cohesion and genre were also included.

Cohesion. As an important discourse feature, cohesion refers to how the events and concepts are connected within a text (McNamara et al., 2012). Among various sources of cohesion, referential cohesion and deep cohesion were two important indices for text complexity and could be assessed by Coh-Metrix (Graesser et al., 2011). Referential cohesion refers to the extent to which there is an overlap or a repetition of words or concepts across sentences, paragraphs, or the entire text, such as replacing "astronaut" with a pronoun "he." For less difficult texts, referential cohesion tended to be lower because the text might cover more topics and use more frequent words (McNamara et al., 2012). Deep cohesion refers to the degree to which the connecting words (e.g., because and therefore) help to clarify the relationship between events, ideas, and information (Graesser et al., 2003; McNamara et al., 2012).

The influence of cohesion on fourth graders' reading comprehension was one part of the research of McNamara et al. (2011). The authors chose four low cohesion texts (two narrative and two expository texts) and revised them to high-cohesion texts by increasing the amount of referential and deep cohesion. For each genre, students were given one low-cohesion text and one high-cohesion text. The result showed that high cohesion texts were easier to understand than low cohesion texts, especially in multiple-choice questions. In addition, students with high knowledge were more beneficial from reading low-cohesion narrative texts than from reading high-cohesion texts. This finding was aligned with previous research (McNamara, 2001; McNamara & Kintsch, 1996; O'Reilly & McNamara, 2007). The authors argued when reading low cohesion texts, students' background knowledge was activated to make inferences and facilitated their filling in of the information gap. However, a reverse effect was also found by O'Reilly and McNamara (2007), that is, high knowledge readers benefited from high-cohesion texts, and low knowledge readers benefited from high-cohesion texts when answering inferential questions. Thus, the effect of cohesion and its interaction with knowledge and question types needs more evidence. In addition, these two studies mentioned above examined the influence of cohesion on reading without differentiation between referential cohesion and deep cohesion. Since the proportion of two kinds of cohesion might be different within a text or across genres, the current study would treat referential cohesion and deep cohesion as two predictors of passage difficulty.

Genre. According to Brooks and Warren (1972), there were four major categories of texts: narrative, expository, description, and persuasion. The current study mainly focused on the narrative and expository due to the test properties of GMRT-4 and WIAT-III. Narrative texts are different from expository texts not only in text features but also in their topics or purposes. In terms of text features, narrative texts tend to have high-frequency words, more difficult sentences, and low cohesion; expository texts are assumed to have difficult words, less challenging sentences, and high cohesion (McNamara et al., 2012). With respect to purpose, narrative texts usually narrate life experiences or express feelings in dialogue with familiar language, while expository texts try to explain concepts or convey new information with unfamiliar terms (Medina & Pilonieta, 2006).

It is generally agreed that narrative texts are less difficult than expository texts. Best et al. (2008) explored the influence of text genre (narrative and expository texts), decoding skills, and word knowledge (i.e., background knowledge) on children's reading comprehension. For each genre, there were three types of questions: free recall task, cued recall task, and multiple choice. Children's decoding skills and world knowledge level were obtained as well. The result showed text genre exerted a notable effect on comprehension. Children's scores on three types of questions were statistically higher in narrative texts than in expository texts, which indicated that narrative text was easier to understand.

For narrative text, decoding skills contributed more than 20% variance in comprehension, much larger than world knowledge. For expository text, world knowledge was the only significant predictor and contributed 14% to 19% variance in three types of questions (Best et al., 2008). The authors argued that for narrative text, students were well familiar with the text structure or schemata; thus, world knowledge was not required as much as decoding skills during the reading process. When comprehending an expository text, children need more world knowledge to help them compensate for their unfamiliarity with the text structure and abstract topics. In addition, lower cohesion in third and fourth grade texts could be another reason for the difficulty in comprehending expository texts. Thus, the passage difficulty across genres could be individually different across students. Notice that Best et al. (2008) only examined the genre effect in WJPC, and the current dissertation would investigate the genre effect on two other standardized tests – GRMT-4 and WIAT-III.

Question Types

As an important facet of test properties, question types may influence test performances due to different processing demand requirements. Basically, two question types were widely investigated in previous research: literal questions and inferential questions. Literal questions usually require students to recall explicitly the information presented in a passage, which is a way of measuring text-level understanding. Inferential questions assess students' inferencing ability which needs students to make inferences by integrating ideas presented in a passage and their prior knowledge (Basaraba et al., 2012; Eason et al., 2012; Hannon, 2012; Miller et al., 2014; Muijselaar et al., 2017).

Different question types may impose different amounts of cognitive demands on readers. Literal questions could require more lower-level skills such as decoding and fluency while inferential questions might need more higher-level skills such as working memory and prior knowledge (Carnine et al., 2010; Herber, 1970; Snider, 1988). To examine whether there was a significant difference in the difficulty of literal, inferential, and evaluative reading items, Basaraba et al. (2012) conducted several analyses on two large independent samples. In their experiment, students were required to read narrative fiction and answer twenty related multiple-choice questions at three points in time: fall, winter, and spring. In the one-way Analyses of Variance (ANOVA), the item difficulty estimated by the Rasch IRT model was the dependent variable, and question types were independent variables. The results showed on average, literal items were significantly easier than inferential and evaluative items. This finding was in line with the levels of comprehension theory which argued literal level comprehension was less challenging than the inferential level (Herber, 1970). Also, the confirmatory factor analysis (Basaraba et al., 2012) further demonstrated that apart from a general reading comprehension factor, there was also a question type effect in reading comprehension performance, which indicated that question type exerted an influence on reading comprehension. This finding was in support of some previous research (e.g., Cain & Oakhill, 2006; Keenan et al., 2008).

However, Muijselaar et al. (2017) argued that literal and inferential question types could not be distinguished in the reading measures of their study. They examined approximately 1,000 fourth graders with two reading comprehension tests. Each test included both narrative and expository texts; a total of seventy-seven items questions were categorized into literal and inferential questions. In spite of the fact that the students' performance on inferential questions was worse than literal questions, the confirmatory factor analysis showed both question types relied on similar reading abilities. The difference was that inferential questions required a higher level of these reading abilities than literal questions. One limitation for this research, as the authors pointed out, was the relatively low interrater reliability of 73% agreement on reading scores, which could impact the results. Therefore, the adoption of standardized tests with a high interrater score might help better understand the relationship between question types and test difficulty.

Impact of Reader-Test Interactions on Test Performance

Under the framework of the RAND model, the reading test performance was the outcome of the interaction between reader characteristics and test properties. Previous studies have found some interaction effects between reader characteristics and passage features, reader characteristics and question types, and passage features and question types. These interactions were reviewed in the following section.

Interactions between Reader and Text

In the study of McNamara et al. (2011), sixty-five fourth graders were recruited, and their decoding skills and prior knowledge were tested. Four low cohesion texts (two narrative and two expository texts) and four revised high cohesion texts were used to test children with three different measures: multiple-choice, cued recall, and free recall. It was found that there was an interaction between readers' prior knowledge and genre. A high level of prior knowledge was more helpful for understanding expository texts than narrative texts, which was similar to the findings in Best et al. (2008). There was also an interaction between cohesion and genre, which meant that students had better performance in the high cohesion narrative texts than in low cohesion texts, but such effect did not occur in expository texts. In addition, the interaction between cohesion and prior knowledge suggested that students with high prior knowledge recalled more information in low cohesion texts. However, they did not find any interaction effect between decoding and cohesion or genre, which indicated the effect of genre and cohesion did not rely on decoding skill.

Kulesz et al. (2016) investigated the interaction between test properties and reader characteristics on GMRT-4. A total of 1,190 students in Grades 7-9 and Grades 10-12 participated in the experiment. In this research, the test properties were explored from six dimensions: word frequency, sentence length, overall passage difficulty, referential cohesion, deep cohesion, and genre. Reader characteristics included word reading/decoding, reading fluency, vocabulary, background/prior knowledge, and working memory. The question types were coded as text memory and text-based inference questions according to their processing demands. EIRM was utilized to detect the interaction effects between test properties and reader characteristics on reading comprehension at the item level.

The results showed that in terms of test properties, genre (expository passages and narrative passages) was a significant predictor of item difficulty for both grade ranges, which was consistent with prior findings (Best et al., 2008; McNamara et al., 2011). Deep cohesion was statistically significant for Grades 10-12. As for reader characteristics, vocabulary and background knowledge had a significant effect on both grade ranges. Working memory was a significant predictor for Grades 7-9, and reading fluency was statistically significant for Grades 10-12.

For Grades 7-9, four interactions were found between: (a) referential cohesion and question types; (b) working memory and deep cohesion; (c) vocabulary and referential cohesion; (d) vocabulary and word frequency. Higher referential cohesion text was easier than lower referential text but had a larger effect on inferential questions, which meant that students had a larger probability of correctly answering inferential questions than literal questions in higher referential cohesion text. For deep cohesion text, students with higher working memory performed better than those with lower working memory regardless of the extent of deep cohesion text. Literal questions were easier in high referential cohesion passages. Students with low vocabulary levels tended to perform worse in low word frequency and low referential cohesion text.

For Grades 10-12, two significant interactions were found between: working memory and question types, and background knowledge and referential cohesion. High working memory helped students answer both text memory and text inference questions. Students with low background knowledge would have more difficulties in reading low cohesion passages, whereas students with high background knowledge could benefit more from low cohesion passages, which was in line with the findings of McNamara et al. (2011). In addition, there was a developmental interaction effect in two grade ranges: Grades 7-9 were dependent on vocabulary while Grades 10-12 were more dependent on background knowledge. Kulesz et al. (2016) concluded that item difficulty was decided by genre and question types, and students' performance on a test was influenced by their developmental differences, test processing demands, or both. However, GMRT-4 was investigated only within middle school and high school students. Given the developmental differences, it would be necessary to investigate the reading performance of elementary students.

Interactions between Reader and Question Types

Miller et al. (2014) utilized crossed random-effect item response models to explore the interplay of passage features, student, and question types. They used 12 expository texts as a baseline model, and the text features of four texts were manipulated to increase the difficulty in cohesion, decoding, syntax, and vocabulary. Reader characteristics included isolated word reading fluency, inferencing, vocabulary, morphological knowledge, and executive functioning (e.g., planning/organization, nonverbal reasoning, working memory). The questions were divided into five types: literal, critical analysis, inferential, comprehension monitoring, and strategy. Comprehension questions and passage fluency (WPM) were outcome measures that were used as proxies for test performance and overall reading competence.

In terms of passage features, results showed the test difficulty of manipulated passages was not statistically different from the original passages, which meant passage difficulty did not influence the probability of correct response. Nonetheless, due to the increased difficulty in decoding and vocabulary in manipulated passages, students' reading speed was slower than the original ones. For reader characteristics, students with higher reading skills had a better performance in the comprehension test. Planning/organization and isolated word reading fluency were significant predictors for passage fluency. With respect to question types, inferential questions and reading strategy questions were more difficult than initial questions; critical analysis and comprehension monitoring questions were not significantly different from initial questions.

There was one significant interaction between planning/organization skill and strategy questions. For students with lower planning/organization skills, the probability of correctly answering strategy questions was similar to that of initial questions. On the other hand, students with higher planning/organization skills did much better in answering literal questions than strategy questions. There was no interaction effect between basic reading skills (e.g., word reading, morphology, and vocabulary) and question types. This result was consistent with the finding of Muijselaar et al. (2017) in which there was a very low correlation between question types and three reading skills (i.e., word reading speed, vocabulary, and working memory). Findings of Miller et al. (2014) and Muijselaar et al. (2017) might contradict the argument that literal questions could require more low-level skills while inferential questions required more high-level skills (Carnine et al., 2010; Herber, 1970; Perfetti et al., 2005). Therefore, whether there was an interaction effect between question types and reader characteristics would need more evidence.

Interactions between Text and Question Types

Eason et al. (2012) investigated the interaction between genre, question types, and reader characteristics. In their research, genre included narrative texts, expository texts, and functional texts (instructional text such as a poster that gives instruction on how to enter a context). Question types were divided into three categories: initial understanding questions (e.g., examining whether students understand the literal meaning of a text); interpretation questions (e.g., emphasizing students' ability to make inferences); critical analysis questions/process strategies questions (e.g., estimating the students' ability to synthesize or evaluate information and required both reading skills and reading strategies). Children aged from 10 to 14 participated in the experiment, and they were tested on various skills: word reading, vocabulary, syntax, listening comprehension, and executive function (e.g., planning and organizing).

The results showed that there was a main effect of genre. Functional texts were easier than narrative and expository texts, but narrative texts were not statistically different from expository texts in text difficulty, which was inconsistent with previous research findings (Best et al., 2008; Kulesz et al., 2016). This could be due to the narrative texts in the test having less referential cohesion than expository texts.

Question types had a significant effect, and there was also an interaction effect between text genre and question types. For functional texts, the performance on the three types of questions was similar to each other; but for narrative and expository texts, initial understanding and interpretation questions were significantly easier than critical analysis questions. It should be noted that the question types interacted with two genres in different ways. For narrative text, initial understanding questions were significantly easier than interpretation questions. For expository text, there was an opposite direction: initial understanding questions were significantly more difficult than interpretation questions.

Eason et al. (2012) argued that in expository text, inference making was happening during the entire reading process so that the interpretation questions might have been answered during reading. But for initial understanding, the same word might appear several times, which could bring difficulties of location for students. When reading narrative texts, students may have been familiar with the schema and did not need more inferences during reading. Thus, the interpretation questions became more difficult.

The impact of the genre in reading comprehension performance also awaited further investigation in the study of Brasher (2017). In her research, three types of maze assessments (i.e., fixed-word deletion test, sentence deletion test, and word-feature deletion test) were constructed to investigate which type of maze tests were validated for informing instruction in fourth grade. Factor analysis showed that all three types of tests measured the same component. However, correlation analysis revealed that the sentence deletion test was the only test type correlated with GMRT-4, and all three types of maze tests were not significantly correlated with WIAT-III. The author pointed out that passage length and genre might be reasons for the differences between the maze tests and validated standardized tests because the three maze tests used longer passages and only adopted expository texts.

In the current dissertation, standardized tests shared one common passage feature, that was, short passages. Therefore, the effect of genre on the test performance could be further investigated. In addition, considering the possible interaction between genre and question type mentioned in Eason et al. (2012), the current study also investigated the effect of such an interaction on test performances, which could also help understand the reason for test differences in Brasher (2017).

Application of EIRM in Reading Test

As a complement of descriptive IRT models, explanatory item response models (EIRM) try to explain item responses with external variables. Variables from both the person side and the item side can make contributions to the understanding and interpretation of test performance. In the following section, the EIRM under the context of the Rasch model, the manipulation of binary data, and its advantages over other methods were reviewed.

Rasch EIRM

Item response models assume that the probability of a correct response to an item can be expressed by a mathematical function between person parameters and item parameters (Dries & Rianne, 2013; Hambleton & Van der Linden, 1982; Thomas, 2011). According to the number of item parameters, there are mainly three IRT models: one-parameter logistic (1PL) model or Rasch model which includes the difficulty parameter (*b*); two-parameter logistic (2PL) model which includes the difficulty parameter (*b*) and the item discrimination parameter (*a*); three-parameter logistic (3PL) model which adds the pseudo-guessing parameter (*c*) to 2PL model (DiTrapani et al., 2018; Kim & Nicewander, 1993; Thomas, 2011). The current study only deals with the Rasch model due to a small sample size because 1PL estimates can be more trustworthy with around 100 respondents (Edelen & Reeve, 2007; Linacre, 1994). For 2PL and 3PL models, reliable parameter estimates may need hundreds or even thousands of respondents (Hulin et al., 1982).

Under the Rasch model, a latent trait or a construct (e.g., ability, skill, or attitude) is assessed by a test. Item discrimination parameter (*a*) is assumed equal, but item difficulty parameter (b) varies across items. This model is descriptive because it aims to inform the position of the person's ability (θ) on the latent trait continuum and the position of item difficulty (b) on the difficulty continuum. The logistic Rasch model (Birnbaum, 1968) can be expressed as:

$$P_i(\theta) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}} \qquad (1)$$

where $P_i(\theta)$ is the probability of a correct response on i-th item by an ability of θ, and b_i is the difficulty parameter of the *i*-th item.

In EIRM, if Y_{ip} is defined as the response by *p*-th person on *i*-th item, the probability of correct Y_{ip} can be modeled with Rasch model as:

$$log\left(\frac{\pi_{p_i}}{1 - \pi_{Pi}}\right) = \eta_{Pi} \qquad (2)$$

where π_{p_i}, equal to $P_i(\theta)$ in equation (1), is the probability of the correct response of *p*-th person on i-th item. Rasch model was transformed to the odds of success probability, and η_{pi} is the log-odds or logit of the success probability. It is also feasible to convert η_{pi} to probability scale by the formula:

$$\pi_{pi} = \frac{e^{\eta_{pi}}}{1 + e^{\eta_{pi}}} \qquad (3)$$

Equation (3) is a logistic function. Notice that in these transformations, the higher values of η_{pi} indicates higher values of odds and probability of correct response.

The basic formulation of EIRM in the context of the Rasch model can be expressed as:

$$\eta_{pi} = \theta_p + \beta_i \tag{4}$$

where η_{pi} is a linear component of person and item characteristics, θ_p refers to a unidimensional latent trait of a person, and β_i is an item easiness parameter. It should be noted in the EIRM literature, item difficulty parameter was modeled as “easiness” instead of “difficulty,” yet it is the negative of the item easiness parameter, -β_i (De Boeck & Wilson, 2004). Since the two terms are only inversely different, the more familiar term – item difficulty – will be used in the current study. It should be kept in mind that the lower values of item parameters indicate more difficult items while higher values of item parameters indicate easier items (Kulesz et al., 2016). The equation (4) showed that higher values of θ_P and β_i produced a higher probability of correct response. In other words, students with higher person ability with tests of easier item difficulty tend to have more successful test performances.

As a complement of descriptive IRT models, EIRM seeks to explain item response by external variables. Take the individual differences into consideration, θ_P can be further explicated through person characteristics:

$$\theta_{pi} = \sum_{j=0}^{J} \vartheta_j \, Z_{pj} + \varepsilon_p, \tag{5}$$

thus

$$\eta_{pi} = \sum_{j=0}^{J} \vartheta_j \, Z_{pj} \; + \varepsilon_p + \beta_i, \tag{6}$$

where *j* is an index for the person predictors, ϑ_j is the fixed regression weight of person property *j*, Z_{pj} is a person predictor and reflects the value of person *p* on person characteristics *j*, and ε_p is a residual person variance or a random effect, normally distributed with the mean of 0.

Similarly, in terms of item difficulty differences, β_i can be explicated through item features:

$$\beta_i = \sum_{k=0}^{K} \gamma_k X_{ki} + \tau_i \, , \tag{7}$$

so that

$$\eta_{pi} = \theta_p + \sum_{k=0}^{K} \gamma_k X_{ki} + \tau_i, \tag{8}$$

where *k* is an index for the item predictors, γ_k is the fixed regression weight of item characteristic *k*, X_{ki} is an item feature predictor and reflects the value of item *i* on item features *k*, and τ_i is a residual item difficulty variance or a random effect, normally distributed with the mean of 0.

Finally, by combining equation (6) and equation (8), item responses can be explained by person characteristics pertaining to person ability (θ_p) and item features pertaining to item difficulty (β_i):

$$\eta_{pi} = \sum_{j=1}^{J} \vartheta_j Z_{pj} + \varepsilon_p \; + \sum_{k=1}^{K} \gamma_k X_{ik} + \tau_i \tag{9}$$

This model was known as the cross-random effect model because it treated the residual of both person ability (θ_p) and item difficulty (β_i) as random effects (De Boeck, 2008). Models in equations (6), (8), and (9) all fall within the framework of the generalized linear mixed model (GLMM) (e.g., De Boeck et al., 2011; Rijmen et al., 2003; Wilson & De Boeck, 2004).

EIRM for Binary Data

EIRM mainly deals with categorical variables, such as binary or dichotomous variables and polytomous data (De Boeck & Wilson, 2004; Fox, 2005). Binary data have two response categories such as "yes" or "no" and "agree" or "disagree." In reading tests, even though multiple-choice questions usually have more than two categories, the correct response is the focus. Thus, this type of data is treated as binary data as well, with correct answers recorded as "1" and other options as "0." Polytomous data usually have more than two categories of response, such as in free response items, 0 points for wrong answers, 1 point for partial correct answers, and 2 for complete correct responses (DiTrapani et al., 2018). The current study mainly dealt with binary data.

Typically, the test data matrix shows in the wide format, which means one row presents one person's performance, and one column presents one item's score as illustrated in Table 1. For binary data, item difficulty is measured as the proportion of correct responses.

Table 1

Test Data Matrix in Wide Format

ID	Item 1	Item 2	Item 3	*Total scores*
Person 1	1	1	1	3
Person 2	1	0	0	1
Person 3	1	1	0	2
Proportion Correct	1	0.67	0.33	

In fact, the item level data in a test have a multilevel or “nested” structure, which means all item responses are nested within persons. In the context of EIRM, all item responses are nested within person and item design factors. Therefore, all item responses were placed in the level one model, and item and person parameters known as crossed design factors were placed in the level two model. The same data can also be presented in the long-data format as illustrated in Table 2.

Table 2

Test Data Matrix in Long Format

ID	Item	Response	Person Characteristics (Z_j)		Item Features (X_k)	
p	i	Y_{pi}	Z_1 (vocabulary)	Z_2 (fluency)	X_1 (literal question)	X_2 (inferential question)
1	1	1	5	7	1	0
1	2	1	5	7	0	1
1	3	1	5	7	0	1
2	1	1	3	5	1	0
2	2	0	3	5	0	1
2	3	0	3	5	0	1

In the long format as shown in Table 2, each person has multiple rows. In the first two columns, “p” and “i” are the index of person and items; “Y_{pi}” in the third column is the item response. For both person characteristics (Z_j) and item features or passage properties (X_k), Table 2 listed some examples of predictors. The first three rows represent person 1's responses on items 1, 2, and 3. It can be seen that person characteristics are constant across items but vary across persons; item features change with items but remain constant across persons.

Advantages of EIRM to Reading Test

In recent years, EIRM has been applied to the area of reading as a novel approach to address both theoretical and practical issues related to the relationship between reader and test (e.g., Eason et al., 2012; Kim et al., 2016; Kulesz et al., 2016; Miller et al., 2014; Spencer et al., 2018). Compared to traditional statistical methods, EIRM has many advantages. First, it has flexibility in modeling predictors, making it possible to explain item responses with external variables. Thus, it expands the service of tests from the measurement purpose to the explanatory function (De Boeck & Wilson, 2004). Second, EIRM is a more comprehensive model framework because it combines standard IRT models with generalized linear models. In this way, psychometrics and statistics are strongly connected in the reading test estimation (Chen & Chen, 2019). Third, since EIRM could directly model and estimate predictive variables from both the item side and the reader side, it can reduce measurement error and provide more accurate estimates for the cross design than the traditional analysis (Debeer & Janssen, 2013).

To summarize, interactions between test properties and reader characteristics have drawn more attention from researchers. EIRM makes it possible to examine such interactions at the item level. An investigation of the reader-test interaction has both theoretical and practical significance. Theoretically, it would help us better understand the reading comprehension process and make contributions to developing reading comprehension theories (De Boeck & Wilson, 2004). Practically, it would provide a clear picture of the characteristics of the standardized reading comprehension tests and afford implications for teachers and researchers in interpreting students' reading performance (Kulesz et al., 2016).

CHAPTER III: METHOD

The data in the current study were part of the dissertation project on maze test validation of Brasher (2017). Two standardized reading comprehension test scores were used as outcome variables, and five component reading skills scores were utilized as reader characteristics variables. Test property variables (i.e., passage features and question types) were obtained from text analyzers and test manuals. This chapter presented the details of the method in four sections: participants, measures, procedure, and analysis.

Participants

The data analyzed in the present study were test results of 89 fourth graders from a rural school district in the Southeastern US. All fourth grade students from 19 classrooms were invited to participate regardless of gender, ethnicity, disability, or intervention. Initially, a total of 100 fourth graders' test data were available, but only those students who completed all tests were included in the current sample. Consequently, 11 students' test results were removed because of incompleteness in one or several tests. Table 3 presents the demographic information of the participants.

Table 3

Demographic Information for the Total Sample

	Percentage of Total (*N* = 89)
Gender, % male	54%
Ethnicity	
% White	70%
% Black	13%
% Hispanic	16%
% Asian	1%
Special Education Services	
Intellectually Gifted	3%
Other Health Impairment	2%
Specific Learning Disability	6%

Measures

Reading Comprehension

Two standardized reading comprehension test scores were used in the current study: *Gates-MacGinitie Reading Test-4th edition* (GMRT-4) Reading Comprehension subtest and *Wechsler Individual Achievement Test- 3rd Edition* (WIAT-III) Reading Comprehension subtest. GMRT-4 was administered in groups, and WIAT-III was administered individually. The fidelity of the administration and scoring procedures were checked, and all procedures were appropriate (Brasher, 2017).

Gates-MacGinitie Reading Test-4th edition (GMRT-4) Reading Comprehension subtest. GMRT-4 is a norm-referenced test and is used to provide a general assessment of reading achievement ability for individual students. For Grade 4, the test consists of two sections: vocabulary and comprehension. The section of comprehension in Form S was administered to students in the sample, and the test has a 35-minute time limit. There were 11 passages ranging from 65 to 125 words in length following by 48 multiple-choice questions in paper form. Students were provided with instructions and item examples at the beginning of testing. The administration group was no more than 26 students at a time. Cronbach's alpha for the current sample group was .93.

Wechsler Individual Achievement Test -3rd Edition (WIAT-III) Reading Comprehension subtest. This comprehension subtest included two expository texts and one narrative text with the word lengths of 60, 102, and 150, respectively. Students could read aloud or silently with no time limit during the testing. The examiner asked two types of questions: literal and inferential questions. The passage was available when students answered related questions. There were 21 items in total, and students orally responded to questions. According to the protocol guidelines, complete, partial, or incorrect answers corresponded to 2, 1, or 0 points. Cronbach's alpha for the current sample group was .70. The inter-rater agreement for this measure was 94.29% (Brasher, 2017).

Reader Characteristics

Reader characteristics such as word reading and decoding, vocabulary, fluency, and morpho-syntactic knowledge were assessed by different measures. These characteristics were utilized as predictors in both GMRT-4 and WIAT-III model analyses. Reader characteristics were consistent across the two standardized tests.

Word Reading and Decoding. This skill was accessed by *WIAT-III Word Reading* and *Pseudoword Decoding subtests*. *WIAT-III Word Reading subtest* requires students to read aloud a list of words with increasing complexity. It has been reported that the average reliability of the subtest was .97 across all grade levels. For fourth grade, the reliability was .98. As for the *Pseudoword Decoding subtest* from the *WIAT-III*, students were required to read aloud from a list of nonsense words with increasing complexity. For this subtest, the average reliability is .98 for all grade levels. The inter-rater agreement was 97.19% for word reading test and 94.38% for pseudoword decoding test (Brasher, 2017).

Reading Fluency. It was accessed by easycbm.com (**EasyCBM**®, Alonzo et al., 2006). Students were asked to read as much as possible of a passage in one minute. Oral reading fluency was determined by the correct words per minute. The test-retest reliability of easycbm.com ranged from .86 to .96 for fourth grade. The inter-rater agreement for this sample was 98.75% (Brasher, 2017).

Reading Vocabulary. It was accessed by *Woodcock-Johnson Tests of Achievement–IV* (WJ-IV) *Reading Vocabulary subtest* (McGrew et al., 2014). Students were required to do a two-part task. In the first part, students read a word and provide a synonym, while in the second part, students should provide an antonym for the targeted word. No decoding help was provided during the task. The manual reported the median reliability of .88 for students aged 5 to 19 and .89 for students aged 9 to 10 years. The inter-rater agreement for this measure was 98.23% (Brasher, 2017).

Morpho-syntactic Knowledge. It was accessed by the *Test of Morphological Awareness* (Carlisle, 2000). Students were required to fill in the blank of a sentence using the appropriate morphological structure (e.g., adjective form or noun form) of a word. Responses were scored correct or incorrect by the examiner. It was recorded that the inter-rater agreement was 99.64%.

Test Properties

Passage Features. Following Kulesz et al. (2016), four passage features were accessed: word frequency, sentence length, cohesion, and genre. The Lexile Framework for Reading (***Lexile*® Text Analyzer**, Schnick & Knickelbine, 2007) was utilized to compute the overall passage difficulty, mean log word frequency (MLWF), and mean sentence length (MSL). Coh-Metrix Text Common Core Ease and Readability Assessor (Coh-Metrix-TERA; Graesser et al., 2011) was used to access the cohesion of passages, such as referential cohesion and deep cohesion. As for genre, the GMRT-4 test manual has provided the genre category (i.e., narrative, expository, and setting) for each passage. Among the 11 passages, there were five narrative passages, five expository passages, and one setting passage. As the technical report claimed, setting text describes scenes and situations and is a familiar part of narrative text. Since the narrativity index of this setting passage (.76) was higher than .50 by Coh-Metrix-TERA, this passage was coded as a narrative text. The genre classification for the passages in WIAT-III was obtained by the narrativity index in Coh-Metrix-TERA. Following Kulesz et al. (2016), taking 50th percentile as a cutting point, passages higher than 50% narrativity was classified as narrative text while equal to or below 50% narrativity was classified as expository text. In WIAT-III, there were two narrative passages and one expository passage.

Question Types. Questions in WIAT-III and GMRT-4 were coded as literal and inferential. For both tests, the test manuals have provided the question type classification information. The two tests classified questions as literal if students could answer by finding information or a restatement explicitly presented in the passage. On the other hand, inferential questions were those that could not be answered by choosing a restatement explicitly stated in the passage (Breaux, 2010; MacGinitie et al., 2000). In GMRT-4, there were 48 items in total, including 25 literal questions and 23 inferential questions. In WIAT-III, among the 21 items, there were 10 literal questions and 11 inferential questions.

Procedure

First, Internal Review Board (IRB) approval was obtained for the current research topic. Second, since the test data from the data holder only have total scores, item level scores were recorded for the two standardized tests. Third, in terms of WIAT-III, even though there were three categories for the item response in test design, half of the item scores of the current sample were binary in nature, which means the response for some items was either 0 credit or 2 credits. In this case, the item parameters should be estimated under mixed models rather than the Rasch model. However, accurate estimation of mixed models has a large sample size requirement (Kim & Wilson, 2020; Kutscher et al., 2019). Since the interest of the current study was to explain the probability of success responses by person characteristics and test properties, both partial credit and full credit could be treated as correct answers compared to incorrect answers. Given this purpose and the small sample size of the current study, both full credit and partial credit were coded as “1”, and the incorrect answer was coded as “0” for EIRM analysis.

At last, data of test properties (i.e., passage features and question types) were obtained by text analyzer software or through test manuals. The MLWF instead of raw word frequency was applied because the logarithmic transformation makes the distribution of word frequency approximates a normal distribution and linearly fit with word processing time (Graesser et al., 2014). It was also found that MLWF had the highest correlation with passage difficulty (*r* = -.78; Lennon & Burdick, 2004). Cohesion such as referential cohesion and deep cohesion were analyzed by Coh-Metrix-TERA (Graesser et al., 2011). Both cohesion indices were presented with percentile test scores. Genre and question types were coded according to the test manuals and narrativity index of Coh-Metrix-TERA. Narrative text and expository text were coded with 0 and 1, and referential cohesion and deep cohesion coded with 0 and 1, respectively.

Analysis

First, descriptive statistics for two standardized reading comprehension tests, reader characteristics, and test properties were computed. Second, bivariate correlation analysis was done to reading comprehension tests, reader characteristics tests, and test properties by two reading comprehension tests. Third, EIRM was used to examine the effects of reader characteristics, test properties, and their interactions on reading comprehension performances. For each test, four models were established through SAS Studio. Model 0 was an unconditional model or a null model without any predictors to estimate the residual variance on reader ability and item difficulty; Model 1 added reader characteristics into Model 0 without text properties to estimate which reader characteristics were crucial for the reader ability, and the research question 1 was answered; Model 2 added test properties (passage features and question type variables) to estimate which variable best predicted the item difficulty, and research question 2 was addressed by this model; Model 3 added both reader characteristics, test properties, and interactions to see whether there was any interaction effect on test performance, and research question 3 was responded. In all models, continuous variables of reader characteristics and text properties were grand mean centered for meaningful interpretation of estimates. All models were estimated through PROC GLIMMIX in SAS Studio.

CHAPTER IV: RESULTS

This chapter presented the data analysis results in each procedure mentioned in the Method section. First, descriptive statistics of the two reading comprehension tests, reader characteristics, and test properties were displayed. Second, bivariate correlation analysis was conducted among two reading comprehension tests, five reader characteristics, and six test properties. Third, four models were built for GMRT-4 and WIAT-III to examine the effects of reader characteristics, test properties, and their interactions on the probability of correct response in two tests. Comparisons were made between the two tests in each model.

Descriptive Statistics

Table 4 displayed the summary statistics of scores of two reading comprehension tests (GMRT-4 & WIAT-III) and reader characteristics. Since the standard scores of different tests were on different metrics, the average of raw scores, the maximum scores, and the minimum scores of each test were reported. All scores were within acceptable limits of skewness and kurtosis (+/-2) for normal distribution (George & Mallery, 2010). GMRT-4, word reading, pseudoword decoding, and fluency all had relatively large standard deviations due to the considerable difference between the maximum score and minimum score. Students’ test scores in these tests had large variability in the current sample.

Table 4

Descriptive Statistics of Reading Comprehension and Reader Characteristics

Test	n	*Mean*	*SD*	Max	Min
GMRT-4	89	25.12	10.66	45	2
WIAT-III	89	15.64	3.23	21	5
Word reading	89	43.43	11.53	68	10
Pseudoword decoding	89	29.62	10.17	50	5
Vocabulary	89	8.81	2.98	13	2
Fluency	89	134.62	36.29	216	22
Morpho-syntactic knowledge	89	14.21	5.17	24	1

Note: GMRT-4 = Gates-MacGinitie Reading Test, 4th edition, Form S; WIAT-III = Wechsler Individual Achievement Test- 3rd edition Reading Comprehension subtests; Max = Maximum; Min = Minimum.

Table 5 compared the results obtained from the preliminary analysis of the test properties of GMRT-4 and WIAT-III. Levene’s tests indicated that equal variance was not assumed between the two tests in terms of MSL, MLWF, Lexile overall difficulty, and deep cohesion. Significant differences were found between the two tests in MSL, *t*(66.14) = 10.18, *p* < .001, MLWF, *t*(21.59) = 2.38, *p* = .03, and referential cohesion, *t*(67) = 3.26, *p* = .002, with GMRT-4 having higher MSL, MLWF, and referential cohesion than WIAT-III. As previously mentioned, MSL and MLWF were two predictors of text difficulty in Lexile, and higher MSL and lower MLWF led to more difficult texts or higher Lexile measures (Lennon & Burdick, 2004). GMRT-4 was more difficult than WIAT-III in MSL and easier in MLWF, but the Lexile measure indicated no significant difference between the two tests in overall passage difficulty, *t*(24.58) = 1.69, *p* = .10. It was not a surprise because the two tests’ passages were all appropriate for fourth grade student level.

Table 5

Descriptive Statistics with Effect Sizes for Test Properties by Two Reading Comprehension Tests

	GMRT-4	WIAT-III	
Variables	# of items = 48	# of items = 21	Cohen's *d*
MSL, *M(SD)****	13.22 (2.86)	8.33 (1.07)	2.26
MLWF, *M(SD)**	3.60 (.17)	3.30 (.55)	.74
Lexile overall difficulty, *M(SD)*	785.33 (131.74)	656.67 (261.02)	.62
Referential cohesion, *M(SD)***	48 (23)	28 (24)	.85
Deep cohesion, *M(SD)*	42 (20)	43 (30)	-.04
Genre			
% Narrative	55	67	N/A
% Expository	45	33	N/A
Question type			
% Literal	52	48	N/A
% Inferential	48	52	N/A

Note. MSL = mean sentence length; MLWF = mean log word frequency; *M* = mean; *SD* = standard deviation. Significant difference at **p* < .05, ***p* < .01, ****p* < .001.

Correlations

Table 6 presents bivariate correlations among the two reading comprehension tests and five reader characteristics. Moderate to large positive correlations were found among these variables. All of the five reader characteristics were statistically correlated with the two standardized reading tests. For GMRT-4, fluency, *r*(87) = .60, *p* < .001, had the highest correlation than other reader characteristics, and vocabulary, *r*(87) = .50, *p* < .001, came second . For WIAT-III, vocabulary, *r*(87) = .72, *p* < .001, and fluency, *r*(87) = .62, *p* < .001, had stronger correlations over other variables.

Among the five reader characteristics, word reading was highly correlated with pseudoword decoding, *r*(87) = .83, *p* < .01), which was expected as both were measuring decoding skills. Word reading was also highly correlated with fluency, *r*(87) = .77, *p* < .001, which indicates that students with high word reading skills in the current sample tended to have strong fluency skills. Likewise, vocabulary was highly correlated with syntax, *r*(87) = .73, *p* < .001, which coincides with previous research (Nagy et al., 2006).

Table 6

Bivariate Correlations of Reading Comprehension and Reader Characteristics

Variables	1	2	3	4	5	6	7
1. GMRT-4	-						
2. WIAT-III	.46***	-					
3. Word reading	.54***	.58***	-				
4. Pseudoword decoding	.46***	.40***	.83***	-			
5. Vocabulary	.50***	.72***	.69***	.51***	-		
6. Fluency	.60***	.62***	.77***	.58***	.66***	-	
7. Morph-syntactic knowledge	.48***	.60***	.66***	.52***	.73***	.59***	-

Note: GMRT-4 = Gates-MacGinitie Reading Test, 4th edition, Form S; WIAT-III = Wechsler Individual Achievement Test- 3rd edition Reading Comprehension subtests. Significant correlations at ****p* < .001.

The results of the correlation analysis among test properties in the two tests are displayed in Table 7. In GMRT-4, MSL was highly correlated with referential cohesion (*r* = .86, *p* < .001) and moderately correlated with deep cohesion (*r* =.32, *p* = .03). MLWF was found to have moderate correlations with referential cohesion (*r* = .47, *p* = .001) and deep cohesion (*r*= .41, *p* = .004). Deep cohesion and referential cohesion were weakly positively correlated (*r* = .29, *p* = .048). Moderated negative correlations were also found between MSL and question type (*r*= -.35, *p* = .02), referential cohesion and question type (*r* = -.44, *p* = .002), and MLWF and genre (*r* = -.34, *p* = .02).

In WIAT-III, it was apparent from Table 7 that MSL had a very strong negative correlation with MLWF (*r*= -.93, *p* < .001), referential cohesion (*r*= -.80, *p* < .001), and deep cohesion (*r* = -.83, *p* < .001). Deep cohesion had a strong negative correlation with genre (*r* = -.79, *p* < .001) and a very strong correlation with MLWF (*r* = .98, *p* < .001). There was also a strong correlation between MLWF and genre (*r* = .65, *p* = .002). Strong correlations among independent variables can cause multicollinearity, which would reduce the accuracy of estimation of parameters (Weissfeld & Sereika, 1991). In the subsequent EIRM analyses, Model 2 with all test properties for WIAT-III did not converge due to high correlations among independent variables. For the four highly correlated test properties (i.e., MSL, MLWF, referential cohesion, and deep cohesion), referential cohesion and deep cohesion might not have reliable estimates from Coh-Metrix due to the short passage lengths of WIAT-III. MSL and MLWF were chosen to rerun the model, but the model still did not converge. As previously shown in Table 5, there was a great discrepancy between GMRT-4 and WIAT-III in MSL (*d* = 2.26). For comparison purpose, MSL were kept and other three variables (MLWF, referential cohesion, and deep cohesion) were deleted in Model 2 and Model 3 estimation for WIAT-III.

Table 7

Bivariate Correlations of Test Properties in GMRT-4 and WIAT-III

			GMRT-4			
Variables	1	2	3	4	5	6
1. MSL	-					
2. MLWF	.14	-				
3. Referential cohesion	.86***	.47**	-			
4. Deep cohesion	.32*	.41**	.29*	-		
5. Question type	-.35*	.13	-.44**	-.02	-	
6. Genre	.19	-.34*	.29*	-.23	-.32*	-
			WIAT-III			
1. MSL	-					
2. MLWF	-.93***	-				
3. Referential cohesion	-.80***	.52*	-			
4. Deep cohesion	-.83***	.98***	.33	-		
5. Question type	-.21	.28	.04	.30	-	
6. Genre	.32	-.65**	.32	-.79***	-.27	-

Note. MSL= mean sentence length; MLWF= mean log word frequency. Significant correlations at *p < .05, **p < .01, ***p < .001.

EIRM Results

Four models were established to investigate the impacts of reader characteristics, test properties, and their interaction effects on GMRT-4. The same procedure was replicated in WIAT-III for comparison. Model 0 or the null model estimated the residual variance on student reading ability (θ) and the item difficulty (b), which postulated that there was a constant reading ability for every item response and a constant difficulty value for all items. For instance, students respond to each item using the same “ability,” no matter the item is decoding-oriented or vocabulary-oriented. And for each item, there was no difference in difficulty among all items. Both the intercept of reading ability and item difficulties were treated as random effects in the current analysis. Model 1 to Model

3 examined the explanatory power of person and item covariates (i.e., reader characteristics and test properties) on the logit of the probability of correct item responses in reading comprehension.

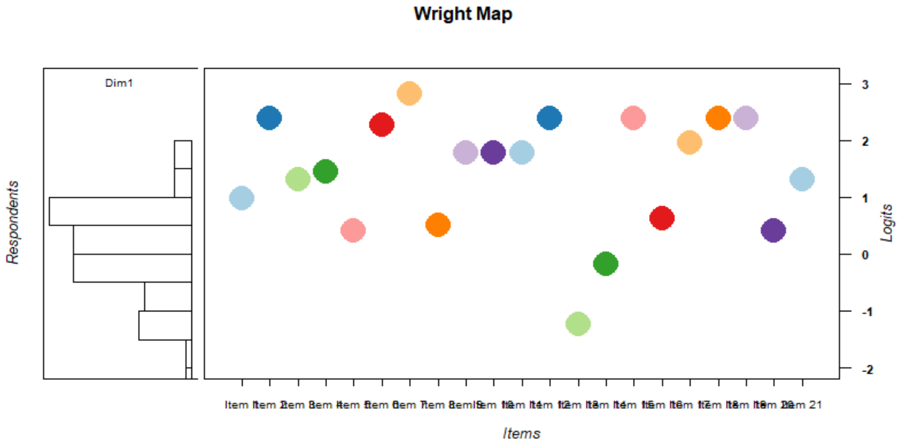
Model Comparisons

Table 8 displays the model fit indices and residual variance of the reader and item side in the four models for GMRT-4 and WIAT-III, respectively. In GMRT-4, the results of Model 0 suggested there was more residual or unexplained variance on the reader side than the item side. This result was expected as reading performance or item response should depend more on students' reading ability than the variability of item difficulties. If the variability of items was more considerable than that of reading ability, test items would not provide reliable estimations on students' reading ability.

However, in WIAT-III, Model 0 indicated minimal residual variance left in the reader side, which meant there was not much variance to be explained by reader characteristics. On the contrary, a larger residual variance was found on the item side, which indicated that the probability of correct responses in WIAT-III relied more on the variability of item difficulty rather than readers' ability. This result could be further explicated by the person-item map in Figure 1. The left panel is the distribution of respondents' or readers' reading ability, and the colored dots are the item difficulty of each item. Of note, the item difficulty level of one third of items (e.g., about seven out of 21 items) is beyond the range of readers' ability, which implies the test items in WIAT-III are too difficult for students in the current sample. In this case, readers' ability contributed limited information to explaining the probability of correct response given the small sample size of both readers and test items. When taking a closer examination of these difficult items, five flagged difficult items (i.e., item 12, 15, 17, 18, and 19) were from the expository texts. There were four items were literal questions (i.e., item 2, 15, 18, and 19) and three items were inferential questions (i.e., item 7, 12, and 17).

Figure 1

WrightMap from the Rasch Model for WIAT-III



Note. This figure shows the distribution of reading ability of 89 students on the left panel and item difficulty of 21items in WIAT-III. The EAP (expected a posteriori) estimates were applied for the ability.

The result of Model 1 showed that for GMRT-4, when reader characteristics or person covariates were added, the reader side residual variance was reduced from 1.41 to .80, which was equal to 43% compared with Model 0. The item side residual variance remained the same as reader characteristics explain students’ reading ability (θ) rather than item difficulty (b). For WIAT-III, it could be seen the residual variance was reduced from .01 to .003 with the addition of reader characteristics. Although the variance reduction was as high as 70%, the explanatory power of reader characteristics was still limited due to the high item difficulty level as shown in Figure 1. The unexplained variance of the item side did not change either.

Table 8

Fit Indices and Residual Variance of Person and Item Side in Estimated Models

GMRT-4

	Fit indices		Reader side		Item side	
	AIC	BIC	Variance (*SE*)	Variance reduction	Variance (*SE*)	Variance reduction
Model 0	4814.69	4808.69	1.41 (.24)	N/A	.88 (.20)	N/A
Model 1	4780.08	4764.08	.80 (.14)	43%	.88 (.20)	0%
Model 2	4803.37	4785.37	1.41 (.24)	0%	.52 (.12)	41%
Model 3	4765.45	4730.61	.80 (.15)	43%	.42 (1.0)	52%

WIAT-III

	Fit indices		Reader side		Item side	
	AIC	BIC	Variance (*SE*)	Variance reduction	Variance (*SE*)	Variance reduction
Model 0	1940.17	1934.17	.01 (.05)	N/A	.72 (.24)	N/A
Model 1	1947.91	1931.91	.003 (.05)	70%	.72 (.24)	0%
Model 2	1941.42	1929.42	.01 (.05)	0%	.55 (.19)	24%
Model 3	1949.39	1925.39	.004 (.05)	60%	.50 (.18)	31%

Note. SE = standard error of variance; Model 0 = without person and item predictors; Model 1= reader characteristics; Model 2 = test properties; Model 3 = reader, test, and reader-test interactions.

Likewise, Model 2 of GMRT-4 exhibited that test property covariates reduced the residual variance of the item side from .88 to .52 compared with Model 0, which was equal to 41% without any influence on the reader side. It was expected because the test properties were regarded as predictors for item difficulty (*b*) rather than students' reading ability (*θ*) in Model 2. For WIAT-III, the test property covariates reduced the residual variance of the item side from .72 to .55, which was equal to 24%. Although there was no significant difference between the overall passage difficulty by Lexile measure (shown in Table 5) between the two tests, the amount of variance reduction from the item side suggested the test properties were more influential on the item difficulty in GMRT-4 than WIAT-III.

As for Model 3 of GMRT-4, when adding reader characteristics, test properties, and their interactive effects together, the residual variance of both the reader side and the item side was reduced compared with Model 0. Notice that the variance reduction on the item side was larger than that on the reader side, which indicated the interaction effects were more influential to item difficulties than the students' reading ability. This reduction pattern was also true for WIAT-III. In other words, the reader-test interaction implied that in both reading tests, the item difficulty of each item was different across readers, or reader characteristics had different impacts across all items.

A likelihood ratio test was applied to test the model fit of the three models. The results showed that for GMRT-4, three models (i.e., Model 1, Model 2, and Model 3) with covariates all statistically better than Model 0, with Model 1, $\Delta\chi^2$ (5) = 44.61, $p < .05$, Model 2, $\Delta\chi^2$ (6) = 23.32, $p < .05$, and Model 3, $\Delta\chi^2$ (15) = 83.27, $p < .05$. In addition, Model 3 was significantly better than Model 1, $\Delta\chi^2$ (10) = 38.66, $p < .05$, and Model 2, $\Delta\chi^2$ (9) = 59.59, $p < .05$, which indicated that Model 3 with interaction effects was better explaining the variance than Model 1 with reader characteristics and Model 2 with test properties. For WIAT, only Model 2 (test property model) was significantly better than Model 0, with $\Delta\chi^2$ (3) = 10.75, $p < .05$. Model 1, $\Delta\chi^2$ (5) = 8.26, $p > .05$, and Model 3, $\Delta\chi^2$ (9) = 14.78, $p > .05$, were not statistically better than Model 0, which indicated that reader characteristics and reader-test interaction effects did not have significant impacts on the test performance of WIAT-III. It was not surprising since there was not much variance to be explained by reader characteristics.

Effects of Reader Characteristics (Model 1)

Table 9 displays parameter estimates of reader characteristics in Model 1. These estimates were analogous to the standardized coefficient like the beta weights in typical regression models. Model 1 for GMRT-4 indicated that fluency ($\beta_{fluency} = .01$, $p < .01$) was statistically significant over other reader characteristics in predicting the logit of the probability of correct item responses. In other words, for fourth graders, students with better fluency skills tended to have a higher likelihood of getting a correct response in GMRT-4. As for WIAT-III, all reader characteristics were nonsignificant, which was expected because the reader covariates did not bring much residual variance reduction, as shown in Table 8. Although the variance reduction percentage was as high as 70%, the reader model was not significantly better than Model 0 due to the small variance in Model 0.

Table 9

Parameter Estimates for Model 1

	GMRT-4			WIAT-III		
Fixed effects	Est.	*SE*	*p*	Est.	*SE*	*p*
Word reading	-.004	.02	.84	-.02	.01	.14
Pseudoword decoding	.01	.02	.42	.01	.01	.38
Vocabulary	.03	.057	.59	.005	.03	.89
Fluency	.01**	.005	.001	.002	.003	.41
Morpho-syntactic Knowledge	.03	.03	.33	.009	.02	.59

Note. Est. = estimates, which show the fixed or random effects in logit, analogous to the standardized regression coefficient (beta weights). *SE* = standard error. Significant effects at ***p* < .01.

Effects of Test Properties (Model 2)

Table 10 presents the parameter estimates of each test property in Model 2 of the two tests. The results of GMRT-4 revealed that MSL (β_{MSL} = -.46, *p* < .001), referential cohesion ($\beta_{referential\ cohesion}$ = 6.68, *p* = .001), deep cohesion ($\beta_{deep\ cohesion}$ = 1.61, *p* = .02), and genre (β_{genre} = -1.06, *p* < .01) were significant predictors of the item difficulty in GMRT-4. Longer sentence length increased the passage difficulty, thus decreasing the chance of getting a correct response. Referential cohesion and deep cohesion were significant positive predictors of the logit of the probability of correct response, which means more referring backward or forward of concepts or connective words increases the chance of getting correct responses. This result was expected as connective words help students figure out the logical relationships across ideas in the passage, making them feel easier to comprehend passages. The significant impact of genre implied that expository passages were more difficult than narrative passages in GMRT-4.

For WIAT-III, as previously mentioned, since MLWF, referential cohesion, and deep cohesion were all highly correlated with MSL, the model estimation did not converge with entering of all of these variables. Therefore, only MSL, genre, and question type were kept for model estimations. The results showed that question type ($\beta_{question\ type}$ = -.76, *p* < .01) was a significant predictor for the probability of a correct response. The negative coefficient meant that inferential questions were more difficult than literal questions in WIAT-III.

Table 10

Parameter Estimates for Model 2

	GMRT-4			WIAT-III		
Fixed effects	Est.	*SE*	*p*	Est.	*SE*	*p*
MSL	-.46***	.13	< .001	-.14	.17	.42
MLWF	-2.63	1.57	.09	-	-	
Referential cohesion	6.68**	2.02	.001	-	-	
Deep cohesion	1.61*	.69	.02	-	-	
Genre	-1.06**	.39	.006	-.42	.39	.29
Question type	-.37	.25	.15	-.76*	.36	.04

Note. Est. = estimates, which shows the fixed or random effects in logit, analogous to the standardized regression coefficient (beta weights). SE = standard error. Significant effects at **p* < .05, ** *p* < .01, ****p* < .001.

Effects of Reader-Test Interactions (Model 3)

Table 11 demonstrates the combined model estimation. As mentioned before, the interaction effects were more influential for the item side compared with the reader side, which indicated that the same item impacted differently across readers due to the varying reader characteristics. For GMRT-4, main effects of fluency ($\beta_{fluency}$ = .02, *SE* = .01, *p* = .001), MSL (β_{MSL} = -.41, *SE* = .13, *p* = .001), referential cohesion ($\beta_{referential\ cohesion}$ = 4.54, *SE* = 2.17, *p* = .03), deep cohesion ($\beta_{deep\ cohesion}$ = 2.18, *SE* = .75, *p* = .004), and genre (β_{genre} = -1.04, *SE* = .46, *p* = .02) were still significant for the probability of getting correct item response after controlling other variables and interactions. Three significant interactions were found: vocabulary and referential cohesion ($\beta_{vocab*refer}$ = -.14, *SE* = .06, *p* < .02), referential cohesion and question type ($\beta_{refer*quest}$ = 2.50, *SE* = 1.21, *p* = .04), and vocabulary and MLWF ($\beta_{vocab*MLWF}$ = -.17, *SE* = .08, *p* < .05). For WIAT-III, since three text covariates (MLWF, referential cohesion, and deep cohesion) were deleted, only one interaction (genre*question type) was examined. The result showed that the interaction between genre and question type was not significant, which was expected as Model 3 was not better than Model 0 according to the model fit test results. Thus, there was no interaction effect in WIAT-III.

Table 11

Parameter Estimates for Model 3

	GMRT-4			WIAT-III		
Fixed effects	Est.	*SE*	*p*	Est.	*SE*	*p*
Word reading	-.004	.02	.84	-.02	.01	.14
Pseudoword decoding	.01	.02	.43	.01	.01	.38
Vocabulary	.03	.06	.58	.005	.03	.89
Fluency	.02**	.01	.001	.002	.003	.41
Syntax	.03	.03	.33	.009	.02	.59
MSL	-.41**	.13	.001	-.16	.17	.33
MLWF	-2.46	1.60	.12	-	-	-
Referential cohesion	4.54*	2.17	.03	-	-	-
Deep cohesion	2.18**	.75	.004	-	-	-
Genre	-1.04*	.46	.02	.26	.62	.67
Question type	-.42	.31	.17	-.03	.64	.97
Vocab*Refer	-.14*	.06	.02	-	-	-
Refer*Quest	2.50*	1.21	.04	-	-	-
Vocab*MLWF	.17*	.08	.048	-	-	-
Genre*Quest	.002	.57	1.00	-1.05	.76	.17

Note. Est. = estimates, which shows the fixed or random effects in logit, analogous to the standardized regression coefficient (beta weights). SE = standard error; Vocab = Vocabulary; Refer = Referential cohesion; Quest = Question type. Significant effects at *$p < .05$, ** $p < .01$.

The three significant interactions in GMRT-4 are graphically depicted in Figure 2 through Figure 5. To plot these three graphs, the low and high levels of reader characteristics and text properties were defined as one standard deviation below or above their respective means. Next, by multiplying the low-level value or high-level value with the log of odds estimates in Table 11, the adjusted logit estimates were obtained for both low-level and high-level reader characteristics or text properties. Then the expected logit of the probability of each interaction effect was obtained. At last, the logit of the probability of correct response was transformed to probability value through formula (10) mentioned in chapter two:

$$\pi_{pi} = \frac{e^{\eta_{pi}}}{1 + e^{\eta_{pi}}} \tag{10}$$

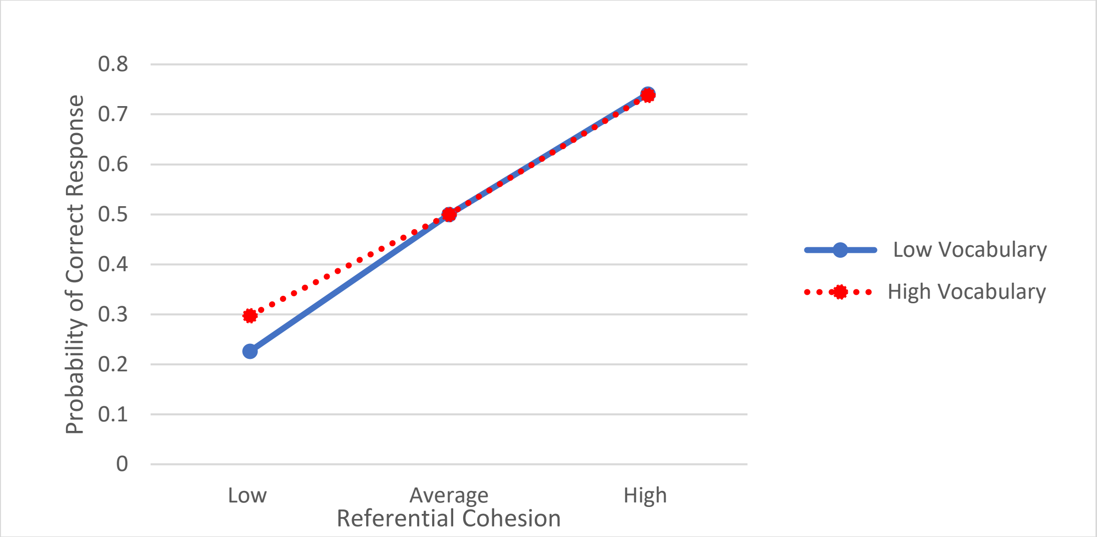
It can also be written as:

$$P\ (Y_{pi}) = \frac{e^{expected\ adjusted\ \log odds}}{1 + e^{expected\ adjusted\ \log odds}} \tag{11}$$

As can be seen in Figure 2, the interaction between vocabulary and referential cohesion reveals that for low referential cohesion passages, students with high vocabulary have a higher probability of getting correct responses than students with low vocabulary. When faced with high referential cohesion passages, the probability of getting a response correct is almost equal for students with high-level and low-level vocabulary. In other words, low cohesion passages were more difficult for students with low vocabulary, but high referential cohesion passages were almost equal to both low and high vocabulary readers. Students with high vocabulary can benefit more from challenging passages with low referential cohesion. Similarly, high referential cohesion passages are helpful for students with low vocabulary to understand.

Figure 2

Interaction between Vocabulary and Referential Cohesion

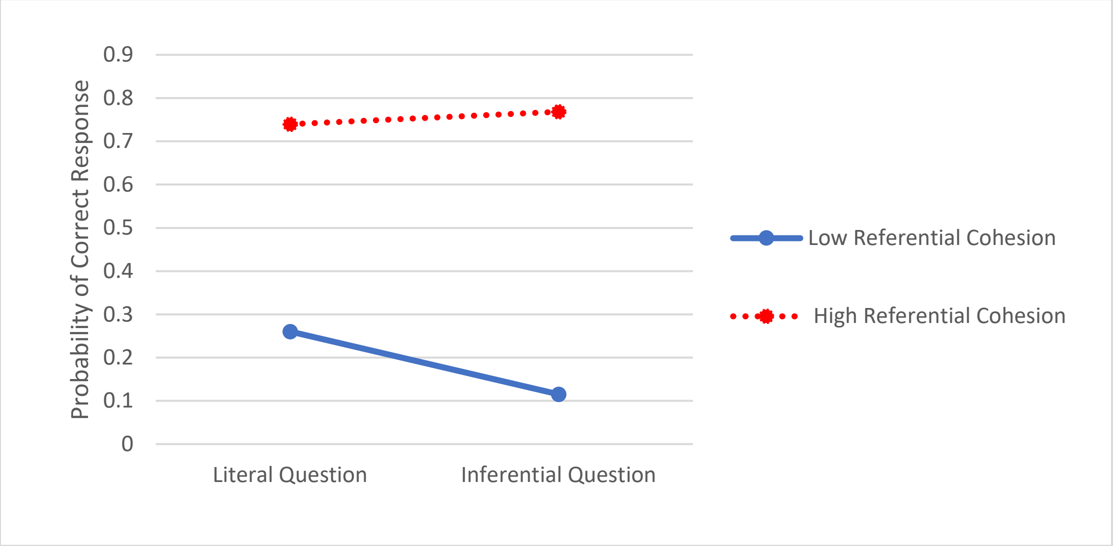


Note. A line plot demonstrating the probability of correct responses for students with low and high vocabulary in face with low and high referential cohesion passages in GMRT-4.

The interaction between referential cohesion and question type is depicted in Figure 3. Passages with high referential cohesion were generally easier than low cohesion passages, with a higher probability of getting correct responses for both literal and inferential questions. For passages with low referential cohesion, literal questions were easier than inferential questions as a higher probability for literal questions was exhibited. However, for high referential passages, the probability of getting a correct response in inferential questions was slightly higher than literal questions, although the difference was small, which is consistent with the findings of Kulesz et al. (2016).

Figure 3

Interaction between Referential Cohesion and Question Type

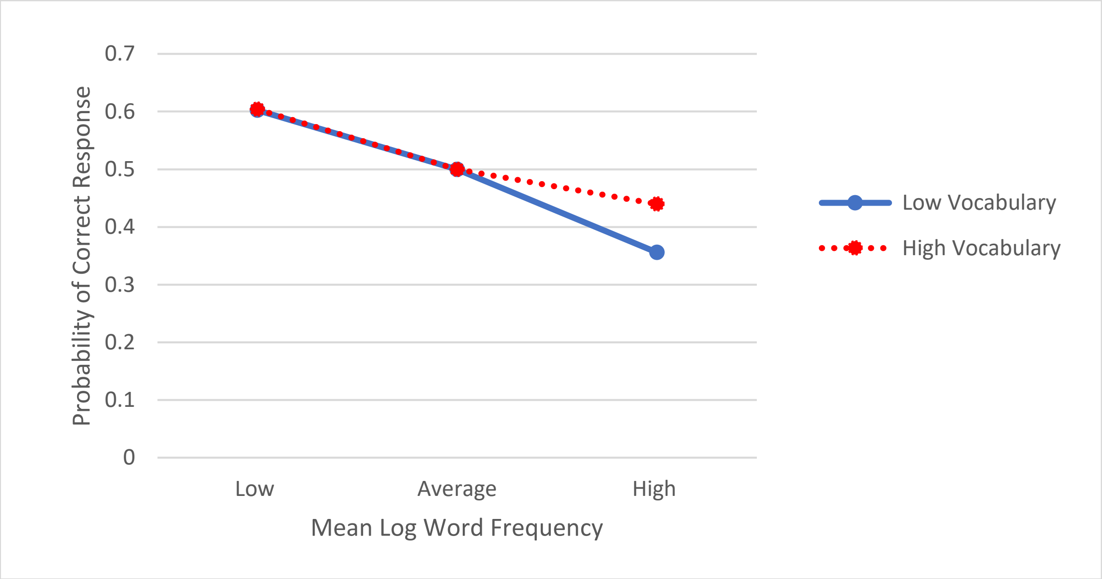


Note. A line plot demonstrating the probability of correct responses for literal and inferential questions for students with low and high vocabulary in GMRT-4.

In Figure 4, there is a clear trend of a decreased probability of correct response as the MLWF increased from low level to high level. For passages with low MLWF, students with high vocabulary had similar performance as students with low vocabulary. On the other hand, for high MLWF passages, high vocabulary students had a higher probability of getting a correct response than low vocabulary students. This was somewhat unexpected since high MLWF usually characterizes less difficult passages. As mentioned before, word frequency in Lexile is not the number of occurrences of a word in a particular text or paragraph, but the frequency that a word occurs in a corpus of 600 million words. It is assumed that an easy passage will have more high-frequent or easy words and a difficult passage will have more difficult or rare words. Recalling both Model 2 and Model 3, the estimates of MLWF were negative values, though not significant. The negative coefficients of MLWF were also reported by Kulesz et al. (2016) which examined the effect of test properties in GMRT-4 for grades 7-9 and grades 10-12, respectively. Therefore, this finding may not be a coincidence. The correlation matrix in Table 7 showed that the correlation between MSL and MLWF was positive, though not significant. It was speculated that in GMRT-4, difficult passages might have higher MLWF.

Figure 4

Interaction between Vocabulary and MLWF



Note. A line plot demonstrating the probability of correct responses for students with low and high vocabulary in face with low and high MLWF in GMRT-4.

Although the three interactions were significant, it should be kept in mind that the variance reduction of Model 3 was small compared with Model 1 and Model 2. As shown previously in Table 8, the percentage of variance reduction for Model 3 was 43% on the reader side and 52% on the item side. Yet, in Model 1 (reader characteristics model), the percentage of variance reduction was 43%, and in Model 2 (test property model) was

41%. That is to say, the interaction effect reduced 11% of residual variance from the item side compared with Model 2 and almost no reduction from the reader side. This is because more variables in the interaction effects came from the test properties (i.e., referential cohesion, question type, and MLWF) than reader characteristics (i.e., vocabulary).

CHAPTER V: DISCUSSION

The initial motivation of the current study was to investigate how reader characteristics, test properties, and their interactions impact the standardized reading comprehension test performance for elementary students within the theoretical framework of the RAND model. Explanatory item response models (EIRM) make it possible to directly model reader side and item side covariates to explain the probability of correct item response or test performance. Although there were studies examining the combined effect of the two sides (i.e., reader and item) on the reading performance, they just focused on one test (Eason et al., 2012; Kulesz et al., 2016). No research has been done so far to compare the reader-test interaction effects between two standardized reading tests so that it is unknown whether the conclusions of one test could be generalized to other tests. The current study systematically investigated the effect of reader characteristics, test properties, and their interactions by ERIM on two standardized tests – GMRT-4 and WIAT-III. By utilizing within-group designs, the similarities and dissimilarities between the two tests were able to be compared. This chapter reviewed the main findings related to the three research questions in chapter one. Limitations for this study and suggestions for future research as well as implications for literacy instruction are discussed.

Reader Characteristics Effects

In the current study, five reader characteristics were examined, including word reading, pseudoword decoding, vocabulary, fluency, and morpho-syntactic knowledge. These reader characteristics were all statistically related to the two standardized reading comprehension tests - GMRT-4 and WIAT-III. The scores of word reading, pseudoword decoding, and fluency had relatively large variability compared to other reader characteristics. Collectively, Model 1 indicated that reader characteristics covariates explained 43% of the variance of reader ability in GMRT-4. Model comparison results showed that Model 1 was better than the null model. However, for WIAT-III, Model 1 was not statistically better than the null model, though the variance reduction percentage of reader characteristics was large. As explained before, since the item difficulty of more than one-third of test items in WIAT-III was beyond the current sample's ability range, not much variance can be explained by reader characteristics. It should be noted there was no significant difference in the overall passage difficulty by Lexile measure between the two tests. Thus, the difficulty level of WIAT-III may come from other sources which would be discussed in the following test property effects section.

Fluency was the only significant predictor of the test performance in GMRT-4 after controlling for other reader characteristics, indicating that students with high fluency can perform better in GMRT-4. In this study, fluency was indicated by the score of words read correctly per minute, which was also indicated as reading rate and accuracy in the study of Sabatini et al. (2018). This finding supports previous research, which suggested fluency was positively related to reading comprehension for fourth graders (Danne et al., 2005; Jenkins et al., 2003; Sabatini et al., 2018). Besides, the significance of fluency in this study adds more evidence for interpreting the performance in GMRT-4. As mentioned earlier, fluency contributed unique variance to the performance of GMRT for upper elementary grades (Kang & Shin, 2019). Of interest, Kulesz et al. (2016) found fluency a significant predictor of GMRT-4 for Grades 10-12, though not easily explained by reading theory. Nevertheless, as mentioned earlier, GMRT-4 has a 35-minute time limit for test-takers, indicating the better test performance may be a requirement for reading speed and accuracy. This requirement is what fluency skills can meet. It can be speculated that students with higher fluency skills tend to have better performance in reading tests with test design and administration requirements similar to those of GMRT-4. Additionally, although vocabulary was not a significant predictor for the reading performance in GMRT-4, it was the only reader characteristics that interact with test properties which would be further discussed in the reader-test interaction section. In the study of Kulesz et al. (2016), vocabulary was a significant predictor of GMRT-4 performance for both grades 7-9 and grades 10-12. The importance of vocabulary in different developmental stages adds more evidence to support vocabulary instruction across grades.

Coincidentally, although reader characteristics did not explain much variance in WIAT-III due to some items that were too difficult, fluency and vocabulary were strongly correlated with WIAT-III, and such correlations were much more robust than with GMRT-4. Kang and Shin (2019) observed that fluency presented a relatively stable influence across reading comprehension tests with elementary school students. It is possible to detect a more accurate relationship between these reader characteristics and WIAT-III after deleting those items that were too difficult in future research.

Test Properties Effects

Through text analyzers and two test manuals, the information of six test properties was obtained, which incorporated MSL, MLWF, referential cohesion, deep cohesion, question type, and genre. Questions were classified into literal and inferential questions, and genres were coded as narrative texts and expository texts. MSL, MLWF, and referential cohesion were statistically different between the two tests, with GMRT-4 having longer sentence length, higher word frequency, and higher referential cohesion than WIAT-III. Nonetheless, there was no significant difference in the overall passage difficulty by Lexile measure between GMRT-4 and WIAT-III.

Since robust correlations among test properties in WIAT-III did not allow the test property model (Model 2) to converge, word frequency, referential cohesion, and deep cohesion were removed from the model. Consequently, while all six test properties remained for GMRT-4, only three test properties (sentence length, genre, and question type) were kept for WIAT-III. The results indicated that overall, test properties explained 41% of the variance in GMRT-4 and 24% of the variance in WIAT-III. The test property models of the two tests were significantly better than the null model.

For GMRT-4, MSL, referential cohesion, deep cohesion, and genre were significant predictors of the test performance. Students had a lower probability of getting correct responses with longer sentence lengths. Higher referential cohesion and deep cohesion made the passages easier to understand. Expository texts were more difficult than narrative texts. For WIAT-III, question type was the only significant predictor.

In terms of cohesion, the positive relationship between the two sources of cohesion (referential cohesion and deep cohesion) and the test performance in GMRT-4 was in agreement with previous findings that high cohesion texts were easier to understand than low cohesion texts (McNamara et al., 2011; O'Reilly & McNamara, 2007). In the current study, referential cohesion and deep cohesion were examined separately, shedding more light on their isolated effects. Unlike Kulesz et al. (2016), which found deep cohesion was the only significant predictor in Grades 10-12, the present study indicated that both the two sources of cohesion were substantial, with referential cohesion more influential than deep cohesion for fourth graders. Remember that referential cohesion emphasizes repetitions of concepts or ideas while deep cohesion helps to clarify the relationships between events or information (Graesser et al., 2003). For fourth graders, students are transiting to the stage of “reading to learn” and facing more challenging texts. Thus, high referential cohesion could be more helpful for students to understand new topics and concepts.

With respect to genre, narrative texts were easier to understand than expository texts in GMRT-4, which is in accord with previous studies (Best et al., 2008). However, for WIAT-III, there was no difference between the two genres. It should be noted that there were 11 passages in GMRT-4 but only three passages in WIAT-III. The limited passage amount of WIAT-III may not provide enough information for differentiating the influences of the two genres.

Question type was the only significant predictor for WIAT-III, indicating that inferential questions were more difficult than literal questions. This result is consistent with the findings of Basaraba et al. (2012). In GMRT-4, the effect of question type was not substantial. As previously discussed, there was no difference in the overall passage difficulty between GMRT-4 and WIAT-III. However, the item difficulty of some items in WIAT-III was beyond the students’ ability. It is speculated that other factors beyond passage features influenced the item difficulty of WIAT-III. The significant effect of the question type in WIAT- III can partly verify this speculation. Of note, one salient difference between GMRT-4 and WIAT-III is the question measure, GMRT-4 using multiple-choice questions while WIAT-III using oral response questions. It is unknown whether the two specific measures cause different effects of question types on the two tests.

Interaction Effects

The combined model (Model 3) examined four reader-test interactions (i.e., vocabulary and referential cohesion, vocabulary and word frequency, referential cohesion and question type, and genre and question type) proposed in research question 3 based on the literature review. Since three variables were removed from WIAT-III model estimation, only one interaction (genre and question type) was investigated and not significant for WIAT-III.

Three significant interactions were found in GMRT-4: vocabulary and referential cohesion, referential cohesion and question type, and vocabulary and word frequency. The results indicated that interaction effects were more influential to item difficulties than reader ability for GMRT-4. For GMRT-4, the interaction effects reduced 11% of the variance from the item side compared Model 3 to Model 2, and almost no variance reduction from the reader side compared Model 3 to Model 1, which is because the sources of interaction effects mainly come from test properties. For WIAT-III, there was 10% and 7% of the variance reduction from the reader side and item side, respectively. Nonetheless, the interaction effect was not significant, and the combined model of WIAT-III was not statistically better than the null model.

These findings suggest that the impacts of the same item or text vary across readers with different reader characteristics or reading component skills. In contrast to the results of Kulesz et al. (2016), who found that interaction effects in GMRT-4 reduced more variance from the reader side rather than the item side for grades 7-9 and grades 10-

12, the current study shows higher variance reduction in the item side than the reader side for fourth grade. One reason is that interaction effects involve more variables coming from test properties rather than reader characteristics. Second, the reader characteristics assessed in the current study are basic reading skills while both basic and high-level component skills were evaluated in the study of Kulesz et al. (2016). Third, the correlations among reader characteristics and GMRT-4 in the current sample were moderate, and some reader characteristics had strong correlations among themselves, which might influence the explanatory power of reader characteristics. Additionally, different from the secondary and high school groups in the study of Kulesz et al. (2016), the current sample is fourth graders or in the elementary stage. Whether there are some developmental differences between reader characteristics and GMRT-4 needs future investigation.

There were two significant reader-text interactions in GMRT-4: vocabulary and referential cohesion, and vocabulary and word frequency. Students with low level vocabulary have a harder time in passages with low referential cohesion and high word frequency passages in GMRT-4. Although vocabulary was not a significant predictor in both the reader model (Model 1) and the combined model (Model 3), the vocabulary level can impact test performance by interacting with test properties (i.e., referential cohesion and MLWF) for fourth graders, which is in line with the findings of Kulesz et al. (2016). As mentioned in the literature review, vocabulary became gradually influential in Grade 4 due to increased expository texts that require a higher demand for vocabulary (Kim, 2020). The interactions between vocabulary and text properties in standardized test results could be one explanation for the findings of Elleman et al.

(2009), which found vocabulary instruction was more effective in comprehending text with custom measures but less effective with standardized measures. It is speculated that the effect of vocabulary is moderated by test properties in standardized tests.

The text-question type interaction in the current study happened to referential cohesion and question type, suggesting that high referential cohesion text is easier to understand than low referential cohesion. Corroborating with previous research (Kulesz et al., 2016), literal questions are less difficult than inferential questions for low cohesion text; but inferential questions are easier than literal questions for high cohesion text. One possible explanation is that high referential cohesion words help students figure out the relations between events and ideas, thus providing benefits for students in making inferences. However, for literal questions, high referential cohesion may bring difficulties for students to locate information.

Limitations of the Study

The present study was limited in several ways. First, the reading passages from GMRT-4 and WIAT-III were relatively short, ranging from 65 to125 and 60 to 165 words, respectively. Although Lexile measure can measure the text complexity with passages less than 200 words, the Coh-Metrix TERA may be less reliable for short passages. In this way, the text properties such as referential cohesion and deep cohesion lack variability, which might influence the model estimation. Nonetheless, in the study of Kulesz et al. (2016), passages for secondary and high school groups were also short and less than 200 words. For standardized tests like GMRT-4 and WIAT-III, given the time limit and students' energy, the number of passages and items always has to balance with each passage's word count. To cover more topics and test items, using shorter passages is a common test design for standardized tests, especially for the K-12 age range. Although short passages in the current study may bring difficulties for test property estimations, it is still meaningful for understanding other tests with similar designs.

The second limitation of the present study is the inherent correlations among passage features and reader characteristics. The correlational analysis showed that all reader characteristics were significantly correlated with each other, and some of the test properties were highly correlated among themselves. Since the passage properties cannot be experimentally controlled, it is hard to examine each factor's unique effect or causal inference on the test performance.

Third, although the current study covers several reader characteristics related to basic reading skills, it is still not exhaustive compared to the range of reader characteristics based on the RAND model. Some high-level reading skills such as working memory, attention, and inference-making are not included. Reader characteristics also involve students' intrinsic and extrinsic motivations which could exert considerable influence on students' test performance (Stutz et al., 2016). As mentioned earlier, there was no interaction between question type and basic reading skills (Miller et al., 2014), but it is unknown whether there are interaction effects between question type and high-level skills. Therefore, a wider array of reader characteristics could be included in a future study.

Last, the relatively small sample size is another limitation of the current study. Although it is acceptable to have around 100 respondents for Rasch EIRM estimation (DiTrapani et al., 2018), the sample size should be increased with the addition of more parameter estimates. Particularly for the test format like WIAT-III, partial credit response may need polytomous item response model estimation, which requires hundreds or even thousands of respondents.

Implications for Practice and Future Research

This study is one of the few studies which systematically investigated the reader-test interactions in different standardized reading comprehension tests. The findings of this study could have important implications for literacy instruction. The results of GMRT-4 are in support of previous findings, highlighting the importance of fluency and vocabulary on reading performances. Higher fluency skills indicate higher reading speed and accuracy, which could meet the requirement of time limits in tests like GMRT-4. The two significant reader-test interactions (i.e., vocabulary and referential cohesion; vocabulary and MLWF) suggest that students with high vocabulary could have a better performance in dealing with more difficult passages. Likewise, in WIAT-III, the strong correlations between WIAT-III and the two reading skills (fluency and vocabulary) indicate the prominent role of these two reader characteristics over other basic reading skills in the upper elementary stage.

Based on these findings, vocabulary and fluency instructions are recommended for literacy instruction in the elementary stage. With the addition of expository texts and more forms of reading materials in upper grades, vocabulary becomes increasingly influential in upper elementary stages and such influence would be long-lasting till middle and high school stages (Kulesz et al., 2016). The intervention of robust vocabulary program may help student lay solid foundation for future reading comprehension development. In terms of fluency, it is not suggested to teaching speed alone but adopted evidence-based fluency approaches to help students achieve both speed and accuracy.

Also, the test properties such as referential cohesion, genre, and deep cohesion are significant test features influencing the item difficulty in GMRT-4, and question types predict the item difficulty in WIAT-III. The reader-test interactions suggested that the impacts of the same passage or item vary across readers with different component reading skills. These findings suggest that educators should pay attention to the test features when choosing standardized tests for reading ability evaluation. For test designers, it is better to isolate the effect of test properties or provide concrete instructions for interpreting how these test properties influence test performances.

The current study may help educators identify critical reading skills and interpret the reasons for poor performance for the two standardized tests. Whether the same conclusions are agreed by other standardized tests needs future examination. Compared with previous test comparison studies (Cutting & Scarborough, 2006; Keenan et al., 2008; Keenan & Meenan, 2014), the application of explanatory item response approach in the current study simultaneously modeling the reader characteristics, test properties, and their interactions, provides a clearer and finer grain size in comparing differences of commonly used standardized tests. It is encouraging to employ the explanatory item response models in more educational research in the future.

REFERENCES

Abbott, R. D., Berninger, V. W., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. *Journal of Educational Psychology, 102*(2), 281-298. https://doi.org/10.1037/a0019318

Adams, M. J. (2001). On the Lexile Framework. In S. White & J. Clement (Eds.), *Assessing the Lexile framework: Results of a panel meeting* (pp. 15-21). National Center for Education Statistics.

Alonzo, J., Tindal, G., Ulmer, K., & Glasgow, A. (2006). *easyCBM online progress monitoring assessment system*. Eugene, OR: Center for Educational Assessment Accountability. Available at http://easycbm.com.

Amendum, S. J., Conradi, K., & Hiebert, E. (2017). Does text complexity matter in the elementary grades? A research synthesis of text difficulty and elementary students' reading fluency and comprehension. *Educational Psychology Review, 30*(1), 121-151. https://doi.org/10.1007/s10648-017-9398-2

Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2012). Examining the structure of reading comprehension: Do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing, 26*(3), 349-379. https://doi.org/10.1007/s11145-012-9372-9

Baumann, J. F. (2009). Vocabulary and reading comprehension: The nexus of meaning. In S.E. Israel, & G.G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 323– 346). Routledge.

Benjamin, R. G. (2011). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review, 24*(1), 63-88. https://doi.org/10.1007/s10648-011-9181-8

Berninger, V. W., Abbott, R. D., Nagy, W., & Carlisle, J. (2010). Growth in phonological, orthographic, and morphological awareness in grades 1 to 6. *Journal of Psycholinguistic Research, 39*(2), 141-163. https://doi.org/10.1007/s10936-009-9130-6

Best, R. M., Floyd, R. G., & McNamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology, 29*(2), 137-164. https://doi.org/10.1080/02702710801963951

Betjemann, R. S., Keenan, J. M., Olson, R. K., & Defries, J. C. (2011). Choice of reading comprehension test influences the outcomes of genetic analyses. *Scientific Studies of Reading, 15*(4), 363-382. https://doi.org/10.1080/10888438.2010.493965

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

Bloom, B. S. (1976). *Human characteristics and school learning*. McGraw-Hill.

Bormuth, J. R. (1969). *Development of readability analysis*. ERIC Clearinghouse.

Brasher, C. F. (2017). *Beyond screening and progress monitoring: An examination of the reliability and concurrent validity of maze comprehension assessments for fourth-grade students* (Publication Number 10266761) [Ph.D., Middle Tennessee State University]. ProQuest Dissertations & Theses Global. Ann Arbor.

Braze, D., Tabor, W., Shankweiler, D. P., & Mencl, W. E. (2007). Speaking up for vocabulary: reading skill differences in young adults. *Journal of Learning Disabilities, 40*(3), 226-243. https://doi.org/10.1177/00222194070400030401

Breaux, K. C. (2010). *Wechsler individual achievement Test-3rd edition (WIAT-III) technical manual with adult norms*. NCS Person, Inc.

Briggs, D. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education, 21*(2), 89-118. https://doi.org/10.1080/08957340801926086

Brooks, C., & Warren, R. P. (1972). *Modern rhetoric*. Harcourt Brace Jovanovich, Inc.

Bryant, P., Nunes, T., & Bindman, M. (2000). The relations between children's linguistic awareness and spelling: The case of the apostrophe. *Reading and Writing, 12*(3), 253-276. https://doi.org/10.1023/A:1008152501105

Burgoyne, K., Whiteley, H., & Hutchinson, J. M. (2011). The development of comprehension and reading‐related skills in children learning English as an additional language and their monolingual, English‐speaking peers. *British Journal of Educational Psychology, 81*(2), 344-354. https://doi.org/10.1348/000709910X504122

Cain, K., & Oakhill, J. (2006). Assessment matters: Issues in the measurement of reading comprehension. *British Journal of Educational Psychology, 76*(4), 697-708. https://doi.org/10.1348/000709905x69807

Cain, K., & Oakhill, J. (2011). Matthew effects in young readers: Reading comprehension and reading experience aid vocabulary development. *Journal of Learning Disabilities*, *44*(5), 431-443. https://doi.org/10.1177/0022219411410042

Cain, K., & Oakhill, J. (2014). Reading comprehension and vocabulary: Is vocabulary more important for some aspects of comprehension? *L'Année psychologique, 114*(04), 647-662. https://doi.org/10.4074/s0003503314004035

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

Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & cognition, 29*(6), 850-859. https://doi.org/10.3758/BF03196414

Carlisle, J. F. (1995). Morphological awareness and early reading achievement. In L. Feldman (Ed.), *Morphological aspects of language processing* (pp. 189–209). Erlbaum.

Carlisle, J. F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing*, *12*(3), 169-190. https://doi.org/10.1023/A:1008131926604

Carlisle, J. F. (2003). Morphology matters in learning to read: A commentary. *Reading Psychology*, *24*(3-4), 291-322. https://doi.org/10.1080/02702710390227369

Carlisle, J. F., & Feldman, L. (1995). Morphological awareness and early reading achievement. In Feldman, L. B. (Ed.), *Morphological aspects of language processing* (pp. 189–209). Erlbaum.

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

Carnine, D. W., Silbert, J., Kame'enui, E. J., & Tarver, S. G. (2010). *Direct instruction reading* (5th ed.). Merrill.

Catts, H. W. (2018). The simple view of reading: Advancements and false impressions. *Remedial and Special Education, 39*(5), 317-323. https://doi.org/10.1177/0741932518767563

Cheatham, J. P., Allor, J. H., & Roberts, J. K. (2014). How does independent practice of multiple-criteria text influence the reading performance and development of second graders? *Learning Disability Quarterly, 37*(1), 3-14. https://doi.org/10.1177/0731948713494016

Chen, P., & Chen, G. (2019). Explanatory item response theory models: Theory and application. *Advances in Psychological Science, 27*(5), 937-950. https://doi.org/10.3724/sp.J.1042.2019.00937

Compton, D. L., Fuchs, D., Fuchs, L. S., Elleman, A. M., & Gilbert, J. K. (2008). Tracking children who fly below the radar: Latent transition modeling of students with late-emerging reading disability. *Learning and Individual Differences, 18*(3), 329-337. https://doi.org/10.1016/j.lindif.2008.04.003

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology, 99*(2), 311-325. https://doi.org/10.1037/0022-0663.99.2.311

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, *10*(3), 277-299. https://doi.org/10.1207/s1532799xssr1003_5

Danne, M. C., Campbell, J. R., Grigg, W. S., Goodman, M. J., & Oranje, A. (2005). *Fourth -grade students reading aloud: NAEP 2002 special study of oral reading*. National Center for Education Statistics, Institution of Education Sciences, U.S. Department of Education.

Davis, F. B. (1944). Fundamental factors of comprehension in reading. *Psychometrika, 9*(3), 185-197. https://doi.org/10.1007/BF02288722

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*(4), 533-559. https://doi.org/10.1007/s11336-008-9092-x

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*, 1–27. http://dx.doi.org/10.18637/jss.v039.i12

De Boeck, P., & Wilson, M. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A Generalized linear and nonlinear approach* (pp. 44–74). Springer. http://dx.doi.org/10.1007/978-1- 4757-3990-9_1

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*(2), 164-185. https://doi.org/10.1111/jedm.12009

Demont, E., & Gombert, J.-E. (1996). Phonological awareness as a predictor of recoding skills and syntactic awareness as a predictor of comprehension skills. *British Journal of Educational Psychology*, *66*(3), 315-332. https://doi.org/10.1111/j.2044-8279.1996.tb01200.x

DiTrapani, J., Rockwood, N., & Jeon, M. (2018). IRT in SPSS using the SPIRIT Macro. *Applied Psychological Measurement*, *42*(2), 173–174. https://doi.org/10.1177/0146621617733956

Dries, D., & Rianne, J. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164-185. https://doi.org/10.1111/jedm.12009

Duke, N. K. (2005). Comprehension of what for what: Comprehension as a nonunitary construct. In S. G. Paris & S. A. Stahl (Eds.), *Center for improvement of early reading achievement (CIERA). Children's reading comprehension and assessment* (pp. 111–122). Lawrence Erlbaum Associates Publishers.

Dunn, L. M., & Markwardt, F. C. (1970). *Peabody individual achievement test*. American Guidance Service.

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader-text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology*, *104*(3), 515-528. http://dx.doi.org/10.1037/a0027182

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*(1), 5-18. https://doi.org/10.1007/s11136-007-9198-0

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness, 2*(1), 1-44. https://doi.org/10.1080/19345740802539200

Elleman, A. M., Steacy, L. M., Olinghouse, N. G., & Compton, D. L. (2017). Examining child and word characteristics in vocabulary learning of struggling readers. *Scientific Studies of Reading, 21*(2), 133-145. https://doi.org/10.1080/10888438.2016.1265970

Fox, J. P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology, 58*(1), 145-172. https://doi.org/10.1348/000711005X38951

Francis, D. J., Fletcher, J. M., Catts, H. W., & Tomblin, J. B. (2005). Dimensions affecting the assessment of reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Center for improvement of early reading achievement (CIERA). Children's reading comprehension and assessment* (pp. 369–394). Lawrence Erlbaum Associates Publishers.

Frantz, R. S., Starr, L. E., & Bailey, A. L. (2015). Syntactic complexity as an aspect of text complexity. *Educational Researcher, 44*(7), 387-393. https://doi.org/10.3102/0013189x15603980

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. https://doi.org/10.1207/S1532799XSSR0503_3

García, J. R., & Cain, K. (2014). Decoding and reading comprehension. *Review of Educational Research, 84*(1), 74-111. https://doi.org/10.3102/0034654313499616

George, D., & Mallery, M. (2010). *SPSS for Windows step by step: A simple guide and reference, 17.0 update.* 10th ed. Allyn & Bacon.

Gernsbacher, M. A. (1997). Two decades of structure building. *Discourse Processes, 23*(3), 265-304. https://doi.org/10.1080/01638539709544994

Gernsbacher, M. A., Varner, K. R., & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 430–445. https://doi.org/10.1037/0278-7393.16.3.430

Geske, A., & Ozola, A. (2009). Different influence of contextual educational factors on boys' and girls' reading achievement. *US-China Education Review*, *6*(4), 38-44.

Goodwin, A. P., & Ahn, S. (2013). A meta-analysis of morphological interventions in English: Effects on literacy outcomes for school-age children. *Scientific Studies of Reading, 17*(4), 257-285. https://doi.org/10.1080/10888438.2012.689791

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*(1), 6-10. https://doi.org/10.1177/074193258600700104

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, *115*(2), 210-229. https://doi.org/10.1086/678293

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223-234. https://doi.org/10.3102/0013189X11413260

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A. P. Sweet, & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82-98). Guilford Press.

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual review of psychology, 48*(1), 163-189. https://doi.org/10.1146/annurev.psych.48.1.163

Hagley, F. (1987). *Suffolk reading scale: Teacher's guide*. NFER-Nelson.

Hambleton, R. K., & van der Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement, 6*(4), 373-378. https://doi.org/10.1177/014662168200600401

Hannon, B. (2012). Understanding the relative contributions of lower-level word processes, higher-level processes, and working memory to reading comprehension Performance in proficient adult readers. *Reading Research Quarterly, 47*(2), 125-152. https://doi.org/10.1002/rrq.013

Hartig, J., Frey, A., Nold, G., & Klieme, E. (2011). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement, 72*(4), 665-686. https://doi.org/10.1177/0013164411430707

Herber, H. L. (1970). *Teaching reading in the content areas*. Prentice Hall.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127-160. https://doi.org/10.1007/BF00401799

Hudson, R. F., Lane, H. B., & Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how? *The Reading Teacher, 58*(8), 702-714. https://doi.org/10.1598/RT.58.8.1

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement, 6*(3), 249-260. https://doi.org/10.1177/014662168200600301

Jenkins, J. R., Fuchs, L. S., Van Den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology, 95*(4), 719-729. https://doi.org/10.1037/0022-0663.95.4.719

Kamil, M.L. (2001). Comments on Lexile framework. In S. White & J. Clement (Eds.), *Assessing the Lexile framework: Results of a panel meeting* (pp. 22-26). National Center for Education Statistics.

Kang, E. Y., & Shin, M. (2019). The contributions of reading fluency and decoding to reading comprehension for struggling readers in the fourth grade. *Reading & Writing Quarterly, 35*(3), 179-192. https://doi.org/10.1080/10573569.2018.1521758

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*(3), 281-300. https://doi.org/10.1080/10888430802132279

Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities*, *47*(2), 125-135. https://doi.org/10.1177/0022219412439326

Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction, 22*(5), 354-367. https://doi.org/10.1016/j.learninstruc.2012.02.001

Kendeou, P., Van Den Broek, P., Helder, A., & Karlsson, J. (2014). A cognitive view of reading comprehension: Implications for reading difficulties. *Learning disabilities research & practice, 29*(1), 10-16. https://doi.org/10.1111/ldrp.12025

Kim, J., & Wilson, M. (2020). Polytomous item explanatory item response theory models. *Educational and Psychological Measurement, 80*(4), 726-755. https://doi.org/10.1177/0013164419892667

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika, 58*(4), 587-599. https://doi.org/10.1007/BF02294829

Kim, Y. S., Wagner, R. K., & Lopez, D. (2012). Developmental relations between reading fluency and reading comprehension: a longitudinal study from Grade 1 to Grade 2. *Journal of Experimental Child Psychology, 113*(1), 93-111. https://doi.org/10.1016/j.jecp.2012.03.002

Kim, Y.-S., Al Otaiba, S., Puranik, C., Folsom, J. S., Greulich, L., & Wagner, R. K. (2011). Componential skills of beginning writing: An exploratory study. *Learning and Individual Differences, 21*(5), 517-525. https://doi.org/10.1016/j.lindif.2011.06.004

Kim, Y.-S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology, 102*(3), 652-667. https://doi.org/10.1037/a0019643

Kim, Y.-S. G. (2017). Multicomponent view of vocabulary acquisition: An investigation with primary grade children. *Journal of Experimental Child Psychology, 162,* 120-133. https://doi.org/10.1016/j.jecp.2017.05.004

Kim, Y.-S. G. (2020). Hierarchical and dynamic relations of language and cognitive skills to reading comprehension: Testing the direct and indirect effects model of reading (DIER). *Journal of Educational Psychology, 112*(4), 667-684. https://doi.org/10.1037/edu0000407

Kim, Y.-S. G., Petscher, Y., & Park, Y. (2016). Examining word factors and child factors for acquisition of conditional sound-spelling consistencies: A longitudinal study. *Scientific Studies of Reading, 20*(4), 265-282. https://doi.org/10.1080/10888438.2016.1162794

Kincaid, J. P., Fishburne Jr, 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.

Kintsch, W., & Kintsch, E. (2005). Comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Center for improvement of early reading achievement (CIERA)*. *Children's reading comprehension and assess*ment (pp. 71–92). Lawrence Erlbaum Associates Publishers.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological review, 85*(5), 363-394. https://doi.org/10.1037/0033-295X.85.5.363

Klare, G. R. (1984). Readability. In P. D. Pearson, R. Barr, & M. L. Kamil (Eds.), *Handbook of Reading Research* (pp. 681–744). Longman.

Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement, 69*(2), 232-244. https://doi.org/10.1177/0013164408322021

Kulesz, P. A., Francis, D. J., Barnes, M. A., & Fletcher, J. M. (2016). The influence of properties of the test and their interactions with reader characteristics on reading comprehension: An explanatory item response study. *Journal of Educational Psychology, 108*(8), 1078-1097. https://doi.org/10.1037/edu0000126

Kuo, L.-j., & Anderson, R. C. (2006). Morphological awareness and learning to read: A cross-language perspective. *Educational psychologist, 41*(3), 161-180. https://doi.org/10.1207/s15326985ep4103_3

Kutscher, T., Eid, M., & Crayen, C. (2019). Sample size requirements for applying mixed polytomous item response models: Results of a Monte Carlo simulation study. *Frontiers in Psychology, 10*, 1-22. https://doi.org/10.3389/fpsyg.2019.02494

Larson, R. K. (2001). Report on the Lexile framework. In S. White & J. Clement (Eds.), *Assessing the Lexile framework: Results of a panel meeting* (pp. 27-33). National Center for Education Statistics.

Lennon, C., & Burdick, H. (2004). The Lexile Framework as an approach for reading measurement and success. Retrieved from http://cdn.lexile.com/cms_page_media/135/The%20Lexile%20Framework%20for%20Reading.pdf

Leslie, L. & Caldwell, J. (2001). *Qualitative reading inventory – 3*. Addison Wesley Longman, Inc.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, *7*(4), 328.

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

MacGinitie, W., MacGinitie, R., Maria, K., & Dreyer, L. (2000). *Gates-MacGinitie reading tests form S & T, Fourth edition*. Riverside.

McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). *Technical manual. Woodcock-Johnson IV*. Riverside.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *55*(1), 51-62. https://doi.org/10.1037/h0087352

McNamara, D. S., Graesser, A., & Louwerse, M. (2012). Sources of text difficulty: Across genres and grades. In J. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 89-116). R & L Education.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*(3), 247-288. https://doi.org/10.1080/01638539609544975

McNamara, D. S., Louwerse, M. M. & Graesser, A. C. (2002). *Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension.* Institute for Intelligent Systems. University of Memphis, Memphis, TN.

McNamara, D. S., Ozuru, Y., & Floyd, R. G. (2011). Comprehension challenges in the fourth grade: The roles of text cohesion, text genre, and readers' prior knowledge. *International Electronic Journal of Elementary Education, 4*(1), 229-257.

Medina, A. L., & Pilonieta, P. (2006). Once upon a time: Comprehending narrative text. In J. S. Schumm (Eds.), *Reading assessment and instruction for all learners* (pp. 222–261). Guilford Press.

Mesmer, H. A., Cunningham, J. W., & Hiebert, E. H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading Research Quarterly, 47*(3), 235-258. https://doi.org/10.1002/rrq.019

Miller, A. C., Davis, N., Gilbert, J. K., Cho, S. J., Toste, J. R., Street, J., & Cutting, L. E. (2014). Novel approaches to examine passage, student, and question effects on reading comprehension. *Learning Disabilities Research & Practice, 29*(1), 25-35. https://doi.org/10.1111/ldrp.12027

Muijselaar, M. M. L., Swart, N. M., Steenbeek-Planting, E. G., Droop, M., Verhoeven, L., & de Jong, P. F. (2017). The dimensions of reading comprehension in Dutch children: Is differentiation by text and question type necessary? *Journal of Educational Psychology, 109*(1), 70-83. https://doi.org/10.1037/edu0000120

Muter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: evidence from a longitudinal study. *Developmental Psychology, 40*(5), 665-681. https://doi.org/10.1037/0012-1649.40.5.665

Nagy, W., Berninger, V. W., & Abbott, R. D. (2006). Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle-school students. *Journal of Educational Psychology, 98*(1), 134-147. https://doi.org/10.1037/0022-0663.98.1.134

Nagy, W. E., Carlisle, J. F., & Goodwin, A. P. (2014). Morphological knowledge and literacy acquisition. *Journal of Learning Disabilities, 47*(1), 3-12. https://doi.org/10.1177/0022219413509967

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

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology, 67*(3), 359-370. https://doi.org/10.1111/j.2044-8279.1997.tb01250.x

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts*. Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers. Retrieved from http://www. corestandards.org/wp-content/uploads/ELA_Standards.pdf.

National Reading Panel. (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*. NIH publication no. 00-4754. Washington, DC: National Institute of Child Health and Human Development. Government Printing Office. Retrieved from http://www.nichd.nih.gov/publications/nrp/report.htm.

Neale, M. (1989). *The Neale Analysis of Reading Ability - Revised*. NFER-Nelson.

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes, 43*(2), 121-152. https://doi.org/10.1080/01638530709336895

Oakhill, J. V., & Cain, K. (2012). The precursors of reading ability in young readers: Evidence from a four-year longitudinal study. *Scientific Studies of Reading, 16*(2), 91-121. https://doi.org/10.1080/10888438.2010.529219

Oakhill, J. V., Cain, K., & Bryant, P. E. (2003). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes*, *18*(4), 443-468. https://doi.org/10.1080/01690960344000008

Oslund, E. L., Clemens, N. H., Simmons, D. C., Smith, S. L., & Simmons, L. E. (2016). How vocabulary knowledge of middle-school students from low socioeconomic backgrounds influences comprehension processes and outcomes. *Learning and Individual Differences*, *45*, 159-165. https://doi.org/https://doi.org/10.1016/j.lindif.2015.11.013

Oslund, E. L., Clemens, N. H., Simmons, D. C., & Simmons, L. E. (2018). The direct and indirect effects of word reading and vocabulary on adolescents' reading comprehension: Comparing struggling and adequate comprehenders. *Reading and Writing*, *31*(2), 355-379. https://doi.org/10.1007/s11145-017-9788-3

Ouellette, G., & Beers, A. (2009). A not-so-simple view of reading: how oral vocabulary and visual-word recognition complicate the story. *Reading and Writing, 23*(2), 189-208. https://doi.org/10.1007/s11145-008-9159-1

Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology*, *98*(3), 554-566. https://doi.org/10.1037/0022-0663.98.3.554

Pearson, P. D. (2000). Reading in the twentieth century. Retrieved from http://www.ciera.org/library/archive/2001-08/0108pdp.pdf.

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.), *Center for improvement of early reading achievement (CIERA)*. *Children's reading comprehension and assessment* (pp. 13–69). Lawrence Erlbaum Associates Publishers.

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

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18*(1), 22-37. https://doi.org/10.1080/10888438.2013.827687

Perfetti, C. A. (1985). *Reading ability*. Oxford university Press.

Perfetti, C. A., & Hart, L. (2001). The lexical basis of comprehension skill. In D. S. Gorfein (Ed.), *Decade of behavior. On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 67–86). American Psychological Association. https://doi.org/10.1037/10459-004

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The Acquisition of Reading Comprehension Skill. In M. J. Snowling & C. Hulme (Eds.), *Blackwell handbooks of developmental psychology*. *The science of reading: A handbook* (pp. 227–247). Blackwell Publishing. https://doi.org/10.1002/9780470757642.ch13

Price, K. W., Meisinger, E. B., Louwerse, M. M., & D'Mello, S. (2015). The contributions of oral and silent reading fluency to reading comprehension. *Reading Psychology, 37*(2), 167-201. https://doi.org/10.1080/02702711.2015.1025118

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8*(2), 185–205. https://doi.org/10.1037/1082-989X.8.2.185

Rimrodt, S., Lightman, A., Roberts, L., Denckla, M. B., & Cutting, L. E. (2005, February). *Are all tests of reading comprehension the same?* [Poster presentation]. The 33rd annual International Neuropsychological Society meeting, St. Louis, MO.

Sabatini, J., Wang, Z., & O'Reilly, T. (2018). Relating reading comprehension to oral reading performance in the NAEP fourth‐grade special study of oral reading. *Reading Research Quarterly, 54*(2), 253-271. https://doi.org/10.1002/rrq.226

Schnick, T., & Knickelbine, M. (2007). *Using the lexile analyzer for educators and media specialists*. Durham, NC: MetaMetrics.

Snider, V. E. (1988). The role of prior knowledge in reading comprehension: A test with LD adolescents. *Direct Instruction News*, 6–11.

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

Spencer, M., Gilmour, A. F., Miller, A. C., Emerson, A. M., Saha, N. M., & Cutting, L. E. (2018). Understanding the influence of text complexity and question type on reading outcomes. *Reading and Writing, 32*(3), 603-637. https://doi.org/10.1007/s11145-018-9883-0

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2006). How accurate are Lexile text measures? *Journal of Applied Measurement, 7*(3), 307-322.

Stutz, F., Schaffner, E., & Schiefele, U. (2016). Relations among reading motivation, reading amount, and reading comprehension in the early elementary grades. *Learning and Individual Differences, 45,* 101-113.

Tannenbaum, K. R., Torgesen, J. K., & Wagner, R. K. (2006). Relationships between word knowledge and reading comprehension in third-grade children. *Scientific Studies of Reading, 10*(4), 381-398. https://doi.org/10.1207/s1532799xssr1004_3

Thomas, M. L. (2011). The value of item response theory in clinical assessment: a review. *Assessment, 18*(3), 291-307. https://doi.org/10.1177/1073191110374797

Thorndike, E. L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychology, 8*(6), 323-332. https://doi.org/10.1037/h0075325

Tong, X., Deacon, S. H., & Cain, K. (2014). Morphological and syntactic awareness in poor comprehenders: another piece of the puzzle. *Journal of Learning Disabilities, 47*(1), 22-33. https://doi.org/10.1177/0022219413509971

Toste, J. R., & Ciullo, S. (2016). Reading and writing instruction in the upper elementary grades. *Intervention in School and Clinic, 52*(5), 259-261. https://doi.org/10.1177/1053451216676835 van den Broek, P., Rapp, D. N., & Kendeou, P. (2005). Integrating memory-based and constructionist processes in accounts of reading comprehension. *Discourse Processes, 39*(2-3), 299-316. https://doi.org/10.1080/0163853X.2005.9651685

van den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1999). The Landscape model of reading: Inferences and the online construction of memory representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 71–98). Lawrence Erlbaum Associates Publishers.

Veenendaal, N. J., Groen, M. A., & Verhoeven, L. (2015). What oral text reading fluency can reveal about reading comprehension. *Journal of Research in Reading, 38*(3), 213-225. https://doi.org/10.1111/1467-9817.12024

Verhoeven, L., & Van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *22*(3), 407-423. https://doi.org/10.1002/acp.1414

Verhoeven, L., van Leeuwe, J., & Vermeer, A. (2011). Vocabulary growth and reading development across the elementary school years. *Scientific Studies of Reading, 15*(1), 8-25. https://doi.org/10.1080/10888438.2011.536125

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2010). *Test of silent reading efficiency and comprehension*. Pro-Ed.

Walczyk, J. J. (2000). The interplay between automatic and control processes in reading. *Reading Research Quarterly, 35*(4), 554-566.

Wanzek, J., Wexler, J., Vaughn, S., & Ciullo, S. (2010). Reading interventions for struggling readers in the upper elementary grades: a synthesis of 20 years of research. *Reading and Writing*, *23*(8), 889-912. https://doi.org/10.1007/s11145-009-9179-5

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological review, 20*(2), 158-177.

Wechsler, D. (2009). *Manual for the Wechsler Individual Achievement Test-(WIAT-III).* Pearson.

Wechsler, D. L. (1992). *Wechsler Individual Achievement Test*. Psychological Corporation.

Weissfeld, L. A., & Sereika, S. M. (1991). A multicollinearity diagnostic for generalized linear models. *Communications in Statistics - Theory and Methods*, *20*(4), 1183–1198. https://doi.org/10.1080/03610929108830558

White, S. (2010). *Understanding adult functional literacy: Connecting text features, task demands, and respondent skills*. Routledge.

Wiederholt, J. L., & Bryant, B. R. (1992). *Gray oral reading tests: GORT-3*. Pro-ed.

Wilson, M., De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models. A generalized linear and nonlinear approach* (pp. 43–74). Springer-Verlag.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of achievement.* Riverside.

Young, A., & Bowers, P. G. (1995). Individual difference and text difficulty determinants of reading fluency and expressiveness. *Journal of Experimental Child Psychology, 60*(3), 428-454. https://doi.org/10.1006/jecp.1995.1048

Yovanoff, P., Duesbery, L., Alonzo, J., & Tindal, G. (2005). Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency. *Educational Measurement: Issues and Practice, 24*(3), 4-12. https://doi.org/10.1111/j.1745-3992.2005.00014.x

Zimmermann, L. M., Reed, D. K., & Aloe, A. M. (2019). A meta-analysis of non-repetitive reading fluency interventions for students with reading difficulties. *Remedial and Special Education,*1-16. https://doi.org/10.1177/0741932519855058