Fall 1-10-2020

An Item-Level and Test-Level Analysis of Struggling Adult Learners' Performance on Reading Assessments

Elena Nightingale
Georgia State University

ACCEPTANCE

This dissertation, AN ITEM-LEVEL AND TEST-LEVEL ANALYSIS OF STRUGGLING ADULT LEARNERS' PERFORMANCE ON READING ASSESSMENTS, by ELENA NIGHTINGALE, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree, Doctor of Philosophy, in the College of Education & Human Development, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chairperson, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty.

____
Daphne Greenberg, Ph.D.
Committee Chair

____
Lee Branum-Martin, Ph.D.
Committee Co-Chair

____
Hongli Li, Ph.D.
Committee Member

____
Ann Kruger, Ph.D.
Committee Member

____
Date

____
Brendan Calandra, Ph.D.
Chairperson, Department of Learning Sciences

____
Paul A. Alberto, Ph.D.
Dean, College of Education & Human Development

AUTHOR'S STATEMENT

____

Elena Nightingale

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Elena Colette Nightingale
1150 Collier Rd. NW
Atlanta, GA. 30318

The director of this dissertation is:

Daphne Greenberg, Ph.D.
Department of Learning Sciences
College of Education and Human Development
Georgia State University
Atlanta, GA. 30303

CURRICULUM VITAE

Elena Nightingale

ADDRESS: 1150 Collier Rd. NW
Atlanta, GA. 30318

EDUCATION:

Ph.D.	2019	Georgia State University Department of Learning Sciences
M.S.	2014	Georgia State University Educational Psychology
B.A.	2012	Wichita State University Secondary Education – ELA/TESOL

PROFESSIONAL EXPERIENCE:

2018-present	Assessment Research & Development Specialist Georgia Department of Education
2012-2018	Graduate Research Assistant Center for the Study of Adult Literacy Georgia State University
2009-2012	Literacy Assistant and Student Teacher Wichita Public Schools

PRESENTATIONS AND PUBLICATIONS:

Nightingale, E., Callaway, C., & Nesbit-McBride, M. (2019, July). Inaugural Administration of the Georgia Alternate Assessment 2.0: Lessons Learned. Paper presented at the 2019 Federal Programs Conference, Atlanta, GA.

Nightingale, E., Branum-Martin, L., & Greenberg, D. (2018, July). A multi-group item-level confirmatory factor analysis of word-reading assessments administered to struggling adult readers." Poster presented at the 28th annual meeting for the Society for Text and Discourse, Brighton, UK.

Nightingale, E., Branum-Martin, L., & Greenberg, D. (2017, July). An IRT analysis of struggling adult readers' performance on orthographic and phonological word reading assessments. Poster presented at the 24th annual meeting of the Society for the Scientific Study of Reading, Halifax, Nova Scotia.

Nightingale, E., Greenberg, D., Branum-Martin, L. (2016). Selecting Fluency Assessments for Adult Learners. *Journal of Research and Practice for Adult Literacy, Secondary, and Basic Education. 5,* 18-29.

Nightingale, E. C., FitzPatrick, E. R., Leroux, A. J., &; Greenberg, D. (2016, April). Using multilevel modeling to identify differential item functioning in a reading assessment given to adults with low literacy skills. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

Nightingale, E., Li, H., & Greenberg, D. (2015, October). A multigroup CFA of the TOSCRF for native and non-native speaking struggling adult learners. Poster presented at the Second Language Research Forum, Atlanta, GA.

Nightingale, E., Branum-Martin, L., & Greenberg, D. (2015, July) Traits versus methods in reading skills among adult learners. In L. Branum-Martin & D. Greenberg (chairs), *Never too late? Unpacking complex literacy skills among adults who struggle with reading*. Paper presented at the 22nd annual meeting of the Society for the Scientific Study of Reading, HI.

Nightingale, E., (2014, April). review of *Improving Adult Literacy Instruction: Options for Practice and Research. Journal of Research and Practice for Adult Literacy, Secondary, and Basic Education.* 2(2), 45-47.

Nightingale, E., Feinberg, I., Brown, J., Bakhtiari, D., Greenberg, D. (2014). *RISE test procedures manual.* Technical Report submitted to The Center for the Study of Adult Literacy, Georgia State University, Atlanta, GA.

Nightingale, E., Greenberg, D., Branum-Martin, L. (2014, March) Measuring reading fluency in adults who struggle with reading. Poster presented at the 4th annual Educational Psychology and Special Education Doctoral Student Association Research Conference, Georgia State University, Atlanta, GA.

Nightingale, E., Greenberg, D., Branum-Martin, L. (2014, Feb.) Measuring silent word reading fluency in adults who struggle with reading. Poster presented at the 37th annual Eastern Educational Research Association (EERA) Conference, Jacksonville, FL.

Nightingale, E., Greenberg, D. (2013, March) Role-focused instruction and demonstrated speaking/listening proficiency in collaborative environments. Poster presented at the 59th Southeastern Psychological Association (SEPA) Annual Conference, Atlanta, GA.

PROFESSIONAL SOCIETIES AND ORGANIZATIONS

2019-present	National Council on Measurement in Education
2018-2019	Society of Text and Discourse
2015-2017	Society for the Scientific Study of Reading
2016-2017	American Educational Research Association

AN ITEM-LEVEL AND TEST-LEVEL ANALYSIS OF STRUGGLING ADULT LEARNERS' PERFORMANCE ON READING ASSESSMENTS

By

ELENA NIGHTINGALE

Under the Direction of Daphne Greenberg and Lee Branum-Martin

ABSTRACT

Selecting and interpreting reading assessments for struggling adult readers can be difficult, as few literacy assessments are designed for this group. In addition, modeling the relations among developing reading skills for adults may be different from what we might expect from a model of reading skills in children. The research here examines the relation between reading skills in struggling adult readers using test-level and item-level models.

The first study models the relations between reading assessments and reading skills in 624 native English speaking adult struggling readers. A series of confirmatory factor analysis (CFA) models were fit based on a multi-trait/multi-method (MTMM) approach. Results from the series of confirmatory factor models indicate that silent word reading accounts for the most trait-related variance in the overall reading model, and speededness accounts for the most method-related variance in the model. The model results reaffirm patterns in past research which indicated that there is a lower correlation/integration of reading skills than found with typically developing children (e.g., Sabatini, Sawaki, Shore, & Scarborough, 2010). It also expands prior research by indicating that specific skills may operate differently in this population and the results can inform our understanding of the overall reading process for struggling adult readers (e.g., Nanda, Greenberg, & Morris, 2014).

The second study is an item-level confirmatory factor analysis of the Test of Irregular Word Reading Efficiency (Reynolds & Kamphaus, 2007) and the Word Attack and Letter-Word Identification subtests of the Woodcock Johnson III Test of Achievement (Woodcock, McGrew & Mather, 2007) which were administered to 931 native and non-native English speaking adults who struggle with reading. Using item-level CFA models structured based on an MTMM approach, this study examines the extent to which the items of these three tests measure general word reading ability versus test-specific skills (traits vs. methods) when administered to adult struggling readers and how that measurement structure may be equivalent in both native and non-native speakers of English. The findings from this study indicate that while group differences are found, structural measurement invariance holds across native and non-native groups under both weak and strong (metric/scalar) measurement invariance conditions.

INDEX WORDS: Adult literacy, Assessment, Multi-Trait Multi-Method Approach, Reading, Confirmatory Factor Analysis, Word Reading

AN ITEM-LEVEL AND TEST-LEVEL ANALYSIS OF STRUGGLING ADULT LEARNERS' PERFORMANCE ON READING ASSESSMENTS

By

Elena Nightingale

A Dissertation

Presented in Partial Fulfillment of Requirements for the

Degree of

Doctor in Philosophy

In

Educational Psychology

In

the Department of Learning Sciences

In

the College of Education and Human Development

Georgia State University

Atlanta, GA

2019

ACKNOWLEDGEMENTS

I would like to thank all of my committee members for their generosity in providing their guidance and expertise in the development of these studies. Dr. Daphne Greenberg has been a wonderful mentor for the past seven years, and her passion and commitment to excellence has inspired this work. My co-chair, Dr. Lee Branum-Martin, has provided guidance at all points of my studies, and has tirelessly supported the technical quality of this research. Not only this research, but my work which has followed and will follow, is immeasurably elevated by exposure to their work ethic and guidance. I am grateful for all the time and effort my committee members, Dr. Hongli Li and Dr. Amy Lederberg, have shared over the past six years in reviewing and supporting this research since the earliest inception, and for also sharing the world of their research with me. An additional thanks is owed to Dr. Ann Kruger, who has joined this research in the final stages, and whose kindness and knowledge as an instructor in my earlier studies has certainly helped shape the views that guide this work. Finally, I am grateful to my family, friends, and spouse, who have been endlessly encouraging as I read many words and wrote many more.

DEDICATION

This work is dedicated to my nieces and nephew, who will also read many words and write many more.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CTOPP	Comprehensive Test of Phonological Processing
GORT	Gray Oral Reading Test
MTMM	Multi-Trait Multi-Method Approach
PAF	Principal Axis Factoring
PPVT	Peabody Picture Vocabulary Test
PTE-A	Pearson Test of English – Academic
RMSEA	Root Mean Square Error of Approximation
SEM	Structural Equation Modeling
SRMR	Standardized Root Mean Square Residual
TIWRE	Test of Irregular Word Reading Efficiency
TLI	Tucker-Lewis Index
TOSCRF	Test of Silent Contextual Reading Fluency
TOSWRF	Test of Silent Word Reading Fluency
TOWRE-PD	Test of Word Reading Efficiency – Phonemic Decoding Efficiency
TOWRE-SW	Test of Word Reading Efficiency – Sight Word Reading Efficiency
WJ-LWI	Woodcock-Johnson - 3rd Edition Letter-Word Identification Subtest
WJ-PC	Woodcock-Johnson - 3rd Edition Passage Comprehension Subtest
WJ-RF	Woodcock-Johnson - 3rd Edition Reading Fluency Subtest
WJ-WA	Woodcock-Johnson - 3rd Edition Word Attack Subtest

A MULTITRAIT-MULTIMETHOD APPROACH TO MODELING THE RELATIONS BETWEEN READINGS SKILLS IN STRUGGLING ADULT READERS

Review of the Literature

When selecting readings tests, it is important to gain as much accurate, interpretable information about the examinee without misinterpreting what the test is measuring, misunderstanding the relation between the skills measured across tests, over-testing, or gathering repetitive information. This selection becomes more complicated with the adult struggling reader population because most literacy measures that assess very elementary skills have not been designed for or normed on an extended age range, particularly with this population. Prior studies have indicated that potential issues may result from administering reading tests designed for children to adults who struggle with reading, as this group has different patterns of strengths and difficulties (e.g. Greenberg, et al., 2011). In addition to issues of assessment, prior research with struggling adult readers has indicated that modeling the relations among developing reading skills may be difficult, and different from what we might expect from a model of reading skills in children (Nanda, Greenberg, & Morris, 2010). To build on this research, this study examines the traits these assessments are measuring in this group, as well as the effect of the method of measurement on assessing reading skills in this group.

Measuring Reading in Struggling Adult Readers

For many of reading assessments, the norming population is only inclusive of children (under 18 years old) or through young adulthood (through approximately 24 years of age); Woodcock, McGrew, & Mather, 2001, Torgesen, Wagner, & Rashotte, 2012, Hammill, Wiederholt, & Allen, 2006, Mather, Hammill, Allen, & Roberts, 2004), adding an additional layer of complexity to how these tests can be used and interpreted with atypical older adult learners. Therefore, not all of the typical reading assessments administered by adult literacy researchers have been normed or been shown to be valid with adults - specifically the population of adults who struggle with reading. In a special issue of the *Journal of Learning Disabilities* focusing on this particular population, Miller, McCardle, and Hernandez (2010) outline the particular social and public health related issues tied to limited literacy skills in a large portion of adults, and explain that an important precursor to addressing limited literacy skill is an accurate assessment of their reading skills.

Studies with reading related assessments administered to adults struggling with reading have found differences when compared to typically developing children. In a study in which the Gray Oral Reading Test (Wiederholt & Bryant, 2001) was administered to adult learners, the results indicated that this assessment did not function with adult learners the way the test developers normed the test on children (Greenberg, Pae, Morris, Calhoon & Nanda, 2009). The authors report that the difference in function could be due to ceiling and basal rules functioning differently for adults than for children, with items not being in the same order of difficulty for adults, as compared to the typically developing child population for which it was designed and normed.

In another study, the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 1997) was administered to struggling adult readers and third-grade children (Pae, Greenberg, & Williams, 2012). The findings indicated that adult learners matched for grade level with typically developing children demonstrated different response patterns. Specifically, when using a sample-dependent approach to measuring item difficulty, items were out of the expected increasing order of difficulty for struggling adult readers (Pae, et al., 2012). This item-level information implies that the vocabulary test used is interacting with a different skill set in the adult population as compared to children, which the authors suggest is likely due to different exposure and experience to words in the two samples, even when controlling for overall reading ability.

A later study by Nanda and colleagues (2014) found that the reliability and validity of the Comprehensive Test of Phonological Processing (CTOPP; Wagner, Torgesen & Rashotte, 1999) dropped significantly from what was reported in the test manual with the norm group of typically developing children when administered to a sample of adult learners (18-72 years old) reading between the third and fifth grade levels. Furthermore, this effect of low reliability and validity increased with age in the adult sample (Nanda, Greenberg, & Morris, 2014). From the studies referenced here, we can see that the issues with measuring reading skills can influence our ability to accurately interpret the results of these reading assessments and others like them.

Modeling Reading in Struggling Adult Readers

Adult struggling readers exhibit deficiencies in all areas related to reading, with a special weakness in phonemic decoding and a relative strength in orthographic decoding (e.g., Eason, Sabatini, Goldberg, Bruce, & Cutting, 2013; Mellard & Fall, 2012; Nanda et al., 2010). In addition to word reading assessments potentially working differently for this population of adult learners, the reading skill sets may operate in different ways for struggling adult readers as compared to typically developing children, resulting in a difference in the way reading skill relation can be modeled in this population. From an initial inspection of the correlation of reading skills, multiple studies with struggling adult readers have found that the correlation of reading skills - including word reading - was lower in struggling adult readers as compared to typically developing children (e.g., Fritz, 2015; Greenberg et al., 2011; Greenberg, Ehri, & Perin, 1997; Mellard & Fall, 2012; Mellard, Fall, and Woods, 2010, Sabatini, Sawaki, Shore, &

Scarborough, 2010). The lower correlations found between these tests indicates a lower integration across skills; that is: these skills are less integrated and may function differently in this group than in the typically developing population of readers. Therefore, it is important to take into consideration that the unique reading patterns and challenges of this population may change the way these tests function, and may change how we can model and interpret them, especially across different developmental ages.

In an analysis of reading comprehension research on struggling adult readers, Tighe and Schatschneider (2014) found that the adults' pattern of lower reading skill correlations resulted in researchers having difficulty applying previously used models to this population. Specifically, an issue identified in this analysis of available research was not only that correlations were lower and models of reading developed for children fit poorly with this population, but also that in general there are several gaps in which components have been examined for this population, particularly in the area of modeling specific subskills and potential moderators or additional factors needed to accurately model reading skills in this population. The authors recommend further examination of these models and the assessments within them before applying to struggling adult readers existing models of reading skills developed with and for typically developing children.

Another approach to modeling reading in struggling adult readers is demonstrated in a study by Mellard et al., (2010) which used path analysis to model the relation of reading skills in adults. The authors found that existing models of the relation between reading skills do not fit well or as intended with this population. They recommend that future researchers examine the relation between reading comprehension, word identification, and phonemic decoding with this population in particular, as these were the most important path factors in these models, and that further examination of the fit of existing theoretical models on this population is needed. A later study by Mellard and colleagues (2015) built on that prior research using Principal Axis Factoring (PAF) to examine the relationship of reading skills in this population with a focus on explaining the variance in overall reading ability by several combinations of specified factors. The PAF approach identifies both strengths and weaknesses, while allowing for analysis of model fit. Results indicated that the factors most useful in predicting overall reading performance in this sample of adult readers were encoding/decoding, vocabulary, processing speed, and working memory. (Mellard, Woods, Desa, & Vuyk, 2015). MacArthur, Konold, Glutting, & Alamprese (2010) approached the question of reliability and construct validity in many of the same measures by analyzing the application of several different specifications of factor models. Their findings indicated that a five-factor model fit best, and that differences based on age were present, with younger populations outperforming older participants in the areas of decoding, spelling, and fluency.

Nanda and colleagues (2010) found that struggling adult readers reading between the 3rd and 5th grade levels demonstrated a low correlation of reading skills, which contributed to the poor fit found for a three, four, and five factor model of overall reading ability. As a result, the authors recommended that future research examine the validity of applying reading measurement approaches developed for children on struggling adult readers (Nanda, et al., 2010). Based on this study, Fritz (2015) compared issues in modeling reading constructs in elementary, middle, and adult readers, fitting multiple factor models from previous research for these age groups, in order to determine if their approach to conceptualizing the relation between reading components needs to vary depending on the age of readers. In older readers, they found weak and even negligible relationships between some reading skills (e.g.: non-word reading and passage comprehension) which otherwise had moderate to strong relationships in elementary and middle school examinees. In terms of model fit for adults, Fritz found that with the increase in age came an increase in factors necessary for adequate model fit, as different reading constructs were not grouping or loading together as cleanly as was found with the elementary and middle school examines.

In conclusion, research indicates that measuring reading and modeling the relation of reading skills in struggling adult readers presents a challenges. One first has to determine whether the measures and models developed for children function as intended with this population. Complexities with applying models of reading abilities with this population need to also be considered.

Multi-Trait/Multi-Method Approach and Assessment Structure

When measuring skills, different test traits and methods may influence performance in different ways (Campbell & Fiske, 1959), and modeling this process can inform our selection and interpretation of tests (Eid & Nussbeck, 2009; Maul, 2013). For the purposes of reading tests in this discussion, method refers to the way in which the test items are administered (e.g., speeded or as part of a similar battery), and trait refers to the construct(s) being assessed (i.e., the cognitive abilities needed to succeed on the items). At least two traits and two methods must be considered in the application of the multi-trait/multi-method (MTMM) approach to interpretation, in order to evaluate the unique contribution of explained variance and to appropriately consider discriminant validity (Platt, 1964). This study applies confirmatory models of MTMM (Eid & Nussbeck, 2009; Maul, 2013) structure to test the roles traits and methods may play for adults who struggle with reading. This approach allows us to analyze the simultaneous effects of assessment methods (e.g., speed, battery) and traits (e.g., phonic decoding, silent reading, and contextual reading) within a selection of reading tests. Moreover, there is the possibility to isolate general traits versus specific traits or sub-abilities (e.g., a specific trait for phonic decoding may simultaneously exist as well as a general ability for reading). By identifying the variance accounted for by the constructs that these tests are intended to measure, as well as that which is accounted for by methods or construct-irrelevant factors, we can better interpret individual and group scores, and can reduce the chances of incorrectly attributing the outcome to test factors irrelevant to the construct measured (Messick, 1989; 1995).

While research using the MTMM approach with reading – particularly within the field of adult literacy – is limited, a couple of studies have used this approach in order to conceptualize the modeling of relation of skills across different reading assessments. In Pae's (2012) study of the Pearson Test of English – Academic (PTE-A; Pearson, 2011), MTMM was the foundational approach for the CFA structure used to analyze the construct validity of this assessment. The results indicated that for these adult English language learners, MTMM revealed a strong trait effect and little to no construct-irrelevant variance to be attributed to a method effect, validating the use of this assessment for this population. The author recommends this approach be used to identify other potential method effects for this and other assessments. In another study, an MTMM framework was used to evaluate the convergent and discriminant validity of reading assessments with struggling adult readers and results indicated a lower correlation of outcomes across measures than expected based on the performance of typically developing children (Nightingale, Greenberg, & Branum-Martin, 2016). Using the MTMM approach in this study, we are able to model the effect of the method of testing and the traits tested, as well as determine the likely structure of reading ability in struggling adult readers.

Purpose of Current Study

The purpose of this exploratory study is to analyze the relations between traits and methods in reading skill assessments administered to native English speaking adult struggling readers, and to discuss the ways in which this may inform the interpretation and use of these tests. This study examines three traits and two methods in a battery of nine reading tests (see Table 1.1). The three traits to be examined are: silent vs. oral reading, contextualized vs. non-contextualized reading, and phonic decoding vs. orthographic word reading. All nine tests measure reading. Three tests in this study measure silent reading, in which an examinee is asked to silently identify words out-of-context, identify words in context, and comprehend sentences. The remaining six tests all require oral reading responses. Three tests were considered measures of contextualized reading if the words presented for the examinees to decode were within the context of a sentence or passage, and the remaining six tests all involved reading words and non-words out of the context of a sentence. Four assessments were categorized as a measure of phonic decoding if participants are asked to decode phonetically-regular non-word blends (nonsense words) or to deconstruct and identify blends of phonetically-regular real words rather than the individual real words found in the other five assessments.

As noted in Table 1.1, the two methods analyzed are speededness (whether a measure being timed impacts outcome), and battery of origin (whether the measures being a part of the same battery impacts outcome). Five of the measures were speeded and four were not. Four of the measures shared a battery – they were a part of the Woodcock Johnson (WJ) battery. Two of the word-reading measures – the TOWRE subtests of sight word reading and phonemic decoding – also shared a battery, but because only two subtests load onto this factor, it cannot modeled in the same way within the factor models. Instead, these subtests were modeled as a residual covariance in the battery model. Two other measures – the TOSCRF and the TOSWRF – were not formally from the same battery, but are created and produced together and represent the same unique method of measuring silent reading fluency. More on how the effect of battery is accounted for in the model sequence is included in the analysis section. The categorization of these assessments for these models, while based on the rationale above, is also in keeping with the skills the tests are designed to measure, as reported in their manuals (Reynolds & Kamphaus, 2007; Hammill et al., 2006; Mather, et al, 2004; Torgesen et al., 2012; Woodcock, et al., 2007).

Table 1.1

Assessment Methods and Traits

	Traits				Methods	
	Reading	Silent	Contextual Reading	Phonic Decoding	Speeded	Battery
TIWRE	x					
TOWRE:PD	x			x	x	TOWRE
TOWRE: SW	x				x	TOWRE
WJ-WA	x			x		WJ
WJ-LWI	x					WJ
WJ-PC	x		x			WJ
WJ-RF	x	x	x		x	WJ
TOSCRF	x	x	x	x	x	TOSWRF/ TOSCRF
TOSWRF	x	x		x	x	TOSWRF/ TOSCRF

Note: TIWRE=Test of Irregular Word Reading Efficiency, TOWRE:PD=Test of Word Reading Efficiency: Phonemic Decoding Efficiency, TOWRE:SW=Test of Word Reading Efficiency: Sight Word Reading Efficiency, WJ=Woodcock Johnson III, WA=Word Attack, LWI= Letter-Word Identification, PC=Passage Comprehension, RF=Reading Fluency, TOSCRF=Test of Silent Contextual Reading Fluency, TOSWRF=Test of Silent Word Reading Fluency

Research Question

This study focuses on the following question: How do reading assessment methods (speed, battery) and traits (silent reading, contextual reading, phonic decoding) in a selection of measures of reading performance relate to each other in a sample of struggling adult readers, and what are the implications for test selection and interpretation for this population? Alternatively, this question can be phrased: What is the structure among methods and traits?

Hypothesis

Based on previous research which indicated that this population of struggling adult readers may have a lower integration of literacy skills, creating complexities and difficulties in modeling overall reading performance (Nanda, et al., 2010), we expected to find a lower correlation among traits (Mellard, Woods, & Fall, 2011; Sabatini et al., 2010) and a higher correlation among methods, which can be referred to as a method effect (Maul, 2013; Sechrest, Davis, Stickle, & McKnight, 2000). Previous research has indicated there may be a pattern of reading skills being more or less difficult for this population, but this does not provide precedent for a more specific hypothesis of which traits or methods may be stronger predictors of overall reading performance (Greenberg et al., 1997). Thus, we expected there may be differences in how these assessments measure what they are designed to measure with this population–which may include method effects or different trait patterns than what might be expected with typical child samples. However, beyond the expectation that a lower correlation of skills would be found, the hypothesis regarding which traits and methods would account for unique variance aside from overall reading ability was largely exploratory for the six models in this study.

Methodology

Participants

Data were collected in the United States and Canada on 624 native English speaking adults who ranged in age from 16 to 73 years old (mean=39, SD=15) and who attended adult literacy programs. Their programs indicated that these adults demonstrated reading skills at the

3rd and 8th grade levels. The majority were female (66%), unemployed (76%), and Black/African-American/African-Canadian/Caribbean (71%). Other ethnic group affiliation included White (25%), Asian (1%), and Native American/Alaska Native/indigenous Canadian (3%). Recruitment of participants occurred at their adult basic education centers, and the data used in this study is a part of data collected on a larger battery of tests which were administered to examinees one-on-one by trained research assistants (IES grant #R305C120001; PI: Daphne Greenberg).

Measures

One measure stands alone (does not come from a battery that includes other tests used in this study):

Irregular Word Reading

The Test of Irregular Word Reading Efficiency (TIWRE; Reynolds & Kamphaus, 2007) assesses irregular word reading efficiency by measuring the examinee's ability to verbally identify phonetically irregular words from a list. The test has been normed on individuals aged 3 to 94 years old and reports internal consistency for all forms in the mid to high .90s. This test is not a speeded measure, and it involves presenting the examinee with phonetically irregular words and letters, which they identify orally until they identify four words incorrectly, after which administration ceases (in this study, all participants started with item number 12).

Two measures come from the Test of Word Reading Efficiency (Torgesen et al., 2012) battery and include:

Sight Word Fluency

The Test of Word Reading Efficiency (TOWRE; Torgesen, et. al., 2012) sight word reading subtest is individually administered and assesses the ability to recognize words which must be orthographically decoded as whole units. The test is normed for examinees 6-24 years old, and has a reported reliability (internal consistency) of .87. Administration for this speeded subtest is 45 seconds, in addition to time required for directions. During the test, the examinee is asked to read aloud from a list of words, while the examiner scores each item as correct or incorrect, from which a final raw score is gathered.

Phonic Decoding Fluency

The Test of Word Reading Efficiency (TOWRE; Torgesen et al., 2012) phonemic decoding subtest is individually administered and assesses the examinee's ability to sound out non-words which must be phonetically decoded to pronounce correctly. The test is normed for examinees 6-24 years old and has a reported reliability (internal consistency) of .87. Administration for this subtest is 45 seconds in addition to the time required for directions. During the test, the examinee is asked to read aloud from a list of non-words, while the examiner scores each item as correct or incorrect, from which a final raw score is gathered.

Four measures come from the Woodcock Johnson III Test of Achievement battery (Woodcock et al., 2007) and include:

Reading Fluency

The Woodcock Johnson III Test of Achievement Reading Fluency subtest (WJ III-RF; Woodcock, et.al, 2007) is designed to assess reading speed by measuring the examinee's ability to silently identify whether a sentence contains correct or incorrect information. The test is normed for examinees ages 2 through 99 years with a reliability (internal consistency) of .90. This speeded test lasts for three minutes, during which the participant is instructed to read each sentence silently and to circle yes or no to identify whether the sentence is correct or incorrect, for as many sentences as they can, within the three minute time limit.

Passage Comprehension

The Woodcock Johnson III Test of Achievement subtest for Passage Comprehension (WJ III-PC; Woodcock et al., 2007) assesses passage comprehension by measuring the examinee's ability to correctly provide the single missing word in a sentence or passage. The measure is normed on individuals aged 2 through 99 years old and is not speeded. The median reliability (internal consistency) reported is .88. Administration involves presenting the examinee with series of sentences with one word left missing and instructing the examinee to read the sentence silently and provide aloud the one word which goes in the blank. All participants started with item 14 and were administered items until six items in a set were answered incorrectly, and moving backward from the starting point of the test as needed until six items in a row are identified correctly.

Phonic Decoding

In the Woodcock Johnson III Test of Achievement Word Attack subtest (WJ III-WA; Woodcock et al., 2007), phonic decoding is assessed by measuring the examinee's ability to pronounce nonsense words. The subtest is normed on individuals from 2 to 99 years of age, at .87 reliability (internal consistency), and is not speeded. The examine is asked to pronounce pseudo-words orally until six words in a row are pronounced incorrectly, and testing backwards as needed until the six lowest items presented are identified correctly (in this study, all participants start with item number 4).

Identification of Letters and Words

The Woodcock Johnson III Test of Achievement subset of Letter-Word Identification (WJ III-LWI; Woodcock et al., 2007) is designed to measure the ability to recognize and orally identify words and letters. This test is not speeded and has been normed on children and adults, ages 2 through 99 years old with internal consistency reliability of .94. The examinee is presented with lists of words, until six words in a row are identified incorrectly, and moving backward from the starting point of the test as needed until six words in a row are identified correctly (in this study, all participants started with item number 33).

Two measures are more informally part of a shared battery, and are developed by the same team, using the same unique method (Mather et al., 2004; Hammill et al., 2006):

Silent Word Reading

The Test of Silent Word Reading Fluency (TOSWRF; Mather et al., 2004) is designed to measure silent reading fluency of single words. It has been normed for examinees aged 6-18 years of age with an average reliability (internal consistency) of .86. This speeded test is 3 minutes in length, preceded by about 1-2 minutes of instruction for a total length of administration of 4-5 minutes. The examinee is presented with lines of words which are printed without spaces, and is asked to draw lines between as many words as possible in 3 minutes.

Silent Contextual Reading

The Test of Silent Contextual Reading Fluency (TOSCRF; Hammill et al., 2006) is intended to measure the speed with which students can silently recognize the individual words in a series of printed passages that become progressively more difficult in content, vocabulary, and grammar. It has been normed on ages 6-18 years old with an average reliability (internal consistency) of .86. This speeded test is 3 minutes in length with a 2 minute practice form and 1-2 minutes of instruction, for a total length of administration of 6-7 minutes. The examinee is presented with passages in which all the words are printed together without spaces and is asked to draw lines between as many correct words as possible in the context of the passage in 3 minutes. (Hammill et al., 2006).

Analytical Approach

This study uses a MTMM approach within a confirmatory factor analysis framework to model the relations between reading constructs, with a focus on the traits and methods involved. Model fit criteria used within this study include the Comparative Fit Index (CFI > .90), Root Mean Square Error of Approximation (RMSEA < .05), and Standardized Root Mean Square Residual (SRMR < .08), and are evaluated based on Barrett (2007) and Hu & Bentler (1999) recommendations for interpreting model fit. Based on the three traits and two methods discussed above, we fit six a priori confirmatory structures, each with its own model (see Table 1.2 and Figures 1.1- 1.6). The sequence of models builds upon the first single-factor model by adding each of the method and trait factors alone, creating a series of bi-factor models. The initial sequence of models that followed included each of the three traits and two methods individually in a bi-factor model with an overall reading factor. This allows for individual interpretation of the variance accounted for by the five factors in question, informing the subsequent composite model. For the trait factors, each was defined as referenced in Table 1, according to the construct being measured by the given subtest. For the method factors, speededness was also defined as categorized above, but the battery factor presented a unique complication, in that one of the sets (TOSCRF/TOSWRF) could not be clearly modeled and interpreted as separate from their shared trait factor, and they were less formally part of the same battery. For this reason, the TOSCRF and TOSWRF, while conceptualized in this study as a shared battery, were not initially included as a covariance in the battery model. The results of all models are discussed in relation to what has previously been found with these skills in this population of struggling adult readers. The test features outlined in Table 1.1 are the basis of the models listed in Table 1.2 which are fit and analyzed in this study.

Table 1.2

Model Sequence

Model	Description
1. Single factor	All scores/performance outcomes are predicted by a single, general factor of reading proficiency
2. Bi-factor, silent word reading	There is an additional specific factor for the silent word reading tests.
3. Bi-factor, contextual word reading	There is an additional specific factor of contextual word reading.
4. Bi-factor, phonic decoding	There is an additional specific factor of phonic decoding.
5. Bi-factor, speededness	There is an additional specific factor of speededness for the five timed tests.
6. Bi-factor, battery	There are additional specific factors for shared battery for six of the subtests.

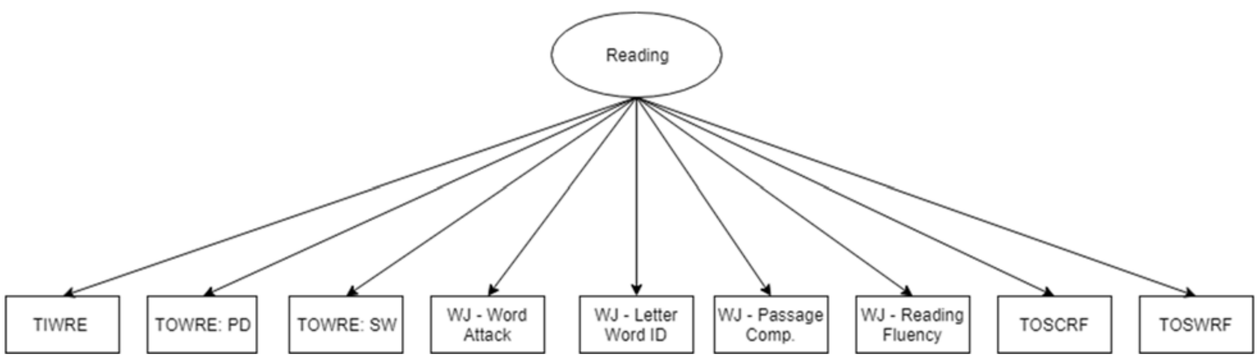

Figure 1.1: Single Factor

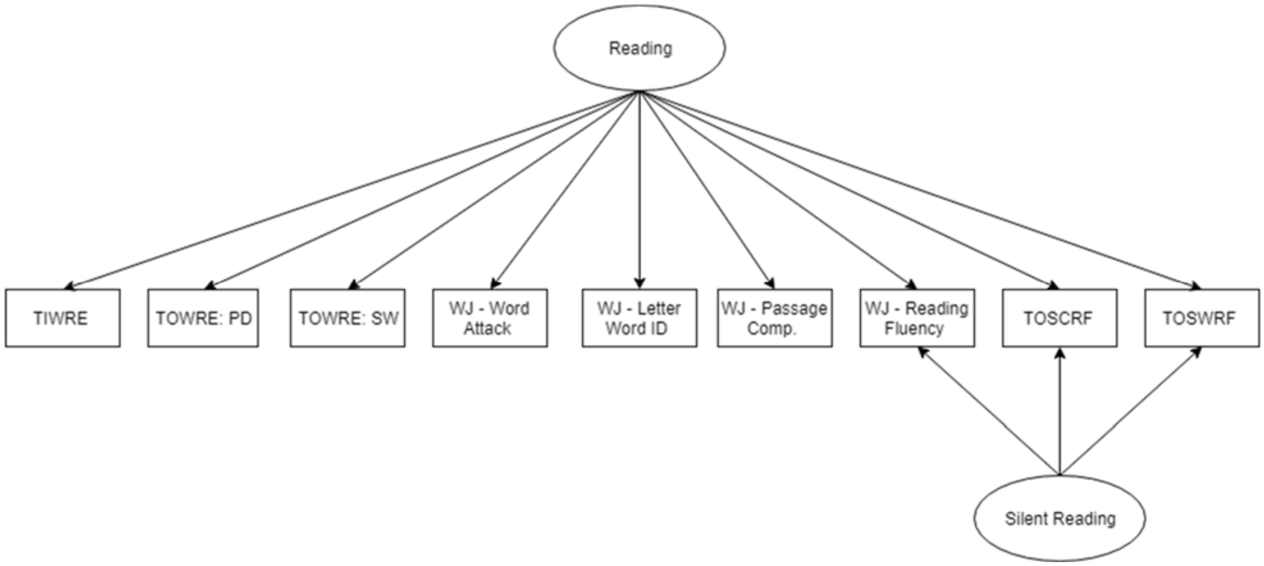

Figure 1.2: Silent Factor

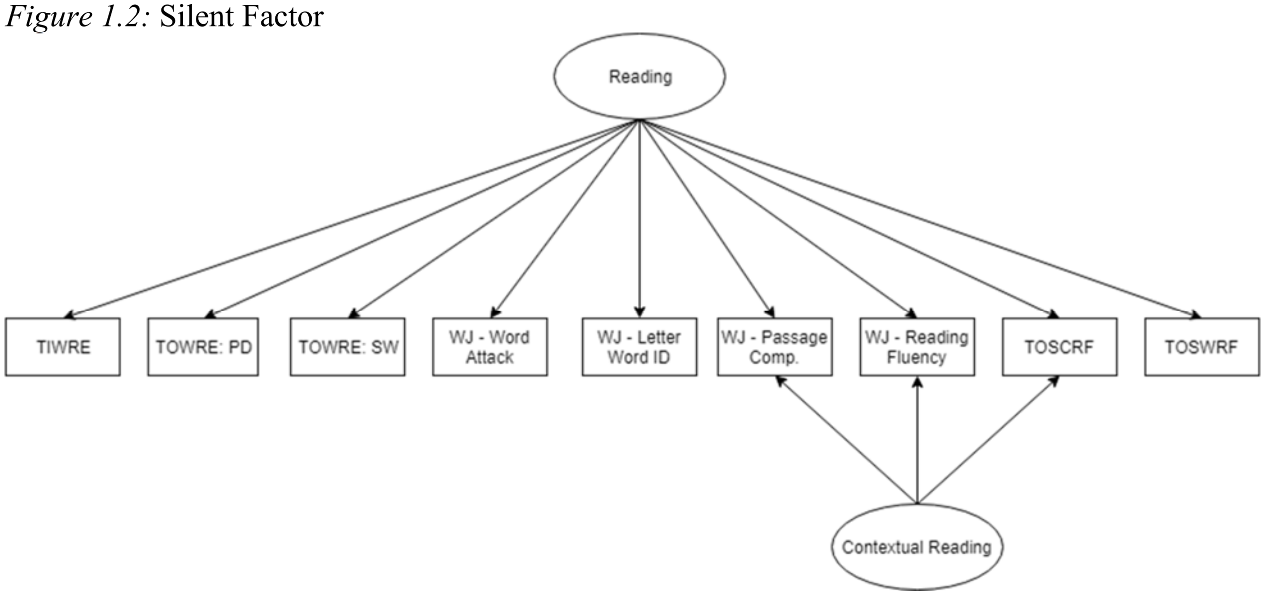

Figure 1.3: Contextual Factor

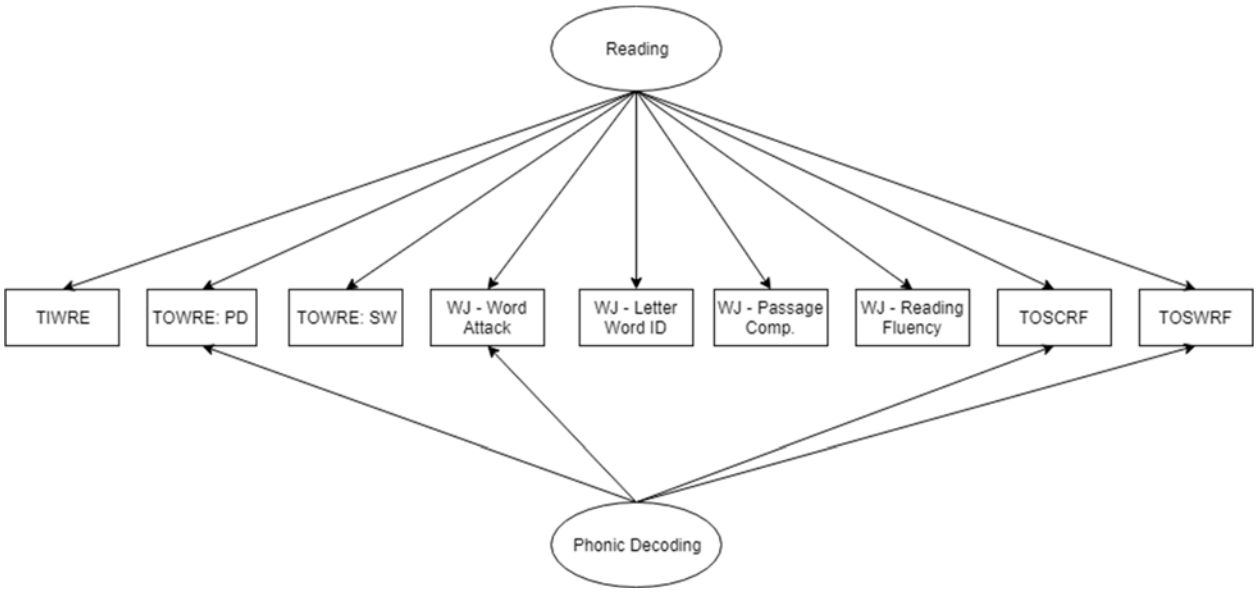

Figure 1.4: Phonic Decoding Factor

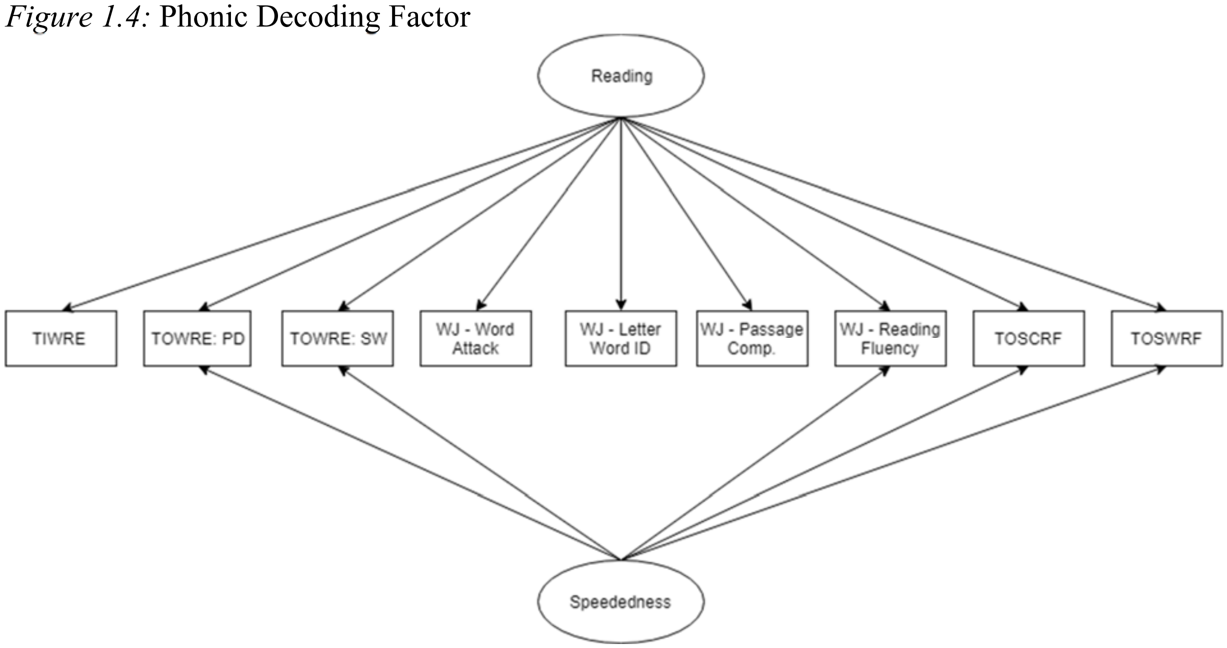

Figure 1.5: Speeded Factor

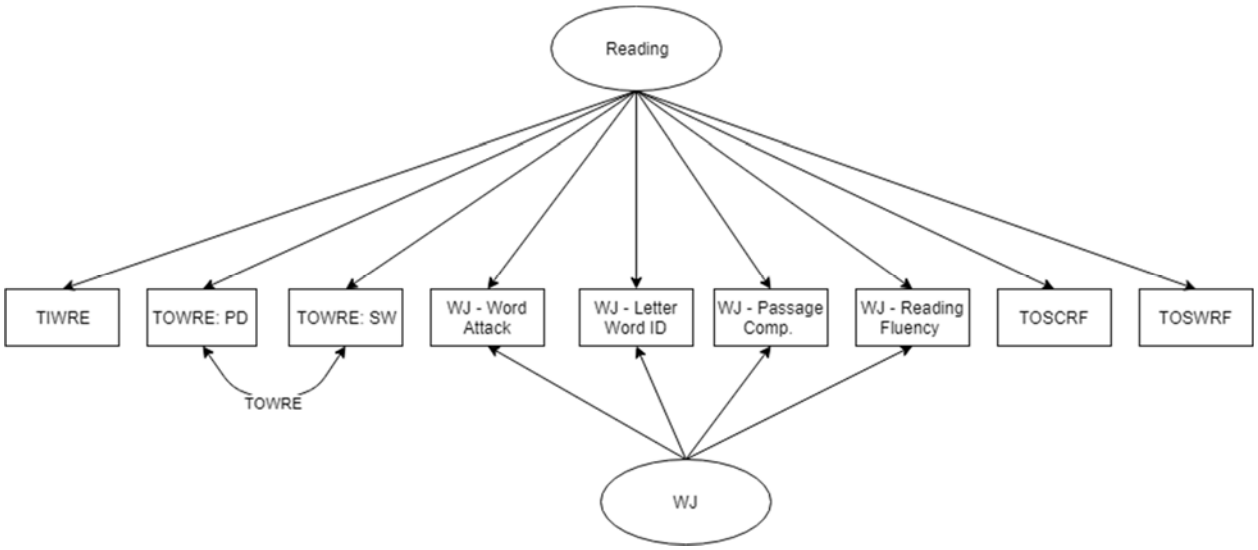

Figure 1.6: Battery Factor

Results

As seen in Table 1.3, participants performed best on two assessments of real word reading (TIWRE and WJ Letter Word ID) and struggled the most on two assessments of non-word phonic decoding (TOWRE Phonemic Decoding Efficiency, WJ Word Attack). The mean grade-equivalents for these tests ranged from 2.2 (TOWRE Phonemic Decoding Efficiency) to 7.1 (TIWRE).

Table 1.3
Performance on Tests

Tests	Raw mean	Raw score min-max	Raw score SD	Standard score mean	GE mean
Test of Irregular Word Reading Efficiency	37.99	20-48	5.07	-	7.1
TOWRE: Phonemic Decoding Efficiency	22.44	0-60	14.83	76.88	2.2
TOWRE: Sight Word Efficiency	66.41	27-108	15.4	80.75	3.5
Test of Silent Contextual Reading Fluency	83.72	9-166	30.53	84.50	5.0
Test of Silent Word Reading Fluency	93.36	2-168	29.39	79.50	5.2
WJIII: Letter-Word Identification	55.00	33-72	8.51	81.96	5.3
WJIII: Passage Comprehension	29.33	16-42	4.41	83.84	4.0
WJIII: Reading Fluency	44.02	7-83	14.03	82.35	5.0
WJIII: Word Attack	16.49	1-31	7.63	79.98	3.1

Note: Standard scores are not available for the Test of Irregular Word Reading Efficiency. TOWRE=Test of Word Reading Efficiency, WJIII=Woodcock Johnson III Test of Achievement/Cognitive Abilities

The nine reading measures in this battery of tests had, as expected, lower correlations with each other than what would be expected in typical populations (see Table 1.4). In general, the lowest correlations were found between the TOSWRF and the WJ Passage Comprehension (r = .380), as well as between the WJ Word Attack and the TOSWRF (r = .404). The highest method correlations were found between tests within the same battery: specifically, between TOSCRF and TOSWRF (r = .803) and the highest trait correlations between tests measuring the same trait were between two phonic decoding tests (WJ Word Attack and TOWRE Phonemic Decoding: r = .830), These ranges of correlations were similar to what was found in previous struggling adult reader studies (Fritz, 2015; Greenberg et al., 2011; Mellard & Fall, 2012; Sabatini et al, 2010).

Table 1.4

Correlations Across Measures

Tests	TIWRE	TOSCRF	TOSWRF	TW-PD	TW-SW	WJ-LWI	WJ-PC	WJ-RF
TIWRE								
TOSCRF	.509*							
TOSWRF	.444*	.803***						
TOWRE: PD	.674**	.483*	.412*					
TOWRE: SW	.603**	.570*	.542*	.696**				
WJ: LWI	.828***	.504*	.436*	.783**	.623**			
WJ: PC	.602**	.435*	.380*	.489*	.519*	.609**		
WJ: RF	.573*	.693**	.604**	.571*	.663**	.569*	.541*	
WJ: WA	.704**	.489*	.404*	.830***	.547*	.789**	.455*	.508*

*Note: TIWRE=Test of Irregular Word Reading Efficiency, TOSCRF=Test of Silent Contextual Reading Fluency, TOSWRF=Test of Silent Word Reading Fluency, TOWRE: PD=Test of Word Reading Efficiency: Phonemic Decoding Efficiency, TOWRE: SW=Test of Word Reading Efficiency: Sight Word Reading Efficiency, WJ=Woodcock Johnson III, LWI= Letter-Word Identification, PC=Passage Comprehension, RF=Reading Fluency, WA=Word Attack. All correlations are significant at p<.0001. *moderate correlation; **strong correlation; ***very strong correlation.*

Single Factor Model

The model fit indices from the single-factor model indicate poor fit when assuming unidimensionality across these assessments for this population (CFI=.779; RMSEA=.193, SRMR=.088) see Table 1.5 for fit indices and the Appendix for loadings), which was not unexpected, considering the breadth of content areas in this battery. The lack of fit indicates that within this population, these reading tests do not measure just one overall factor of reading, but that there are additional factors accounting for the variance in performance.

Bi-Factor Model – Silent Word Reading

To account for the first trait considered here, a bi-factor model was fit which included an additional factor for silent status - that is, the assessments which assessed word or sentence reading silently – the TOSCRF, TOSWRF, and WJ Reading Fluency subtest. This model produced a significant improvement in fit (CFI=.912, RMSEA=.129, SRMR=.052) when compared to the fit of the single-factor model, though there was still room for model fit improvement and further traits and methods to explore in the previously outlined sequence.

Bi-Factor Model – Contextual Word Reading

We fit a model with a factor accounting for contextuality - whether the words are presented in the context of a sentence or out of context as a list of unrelated words. While the addition of the context factor in this model significantly ($p<.001$) improved model fit from the single-factor model, the overall fit indices were not as good as the model with the silent word reading factor (CFI=.814, RMSEA=.184 SRMR=.086). Of note, one of the two cases of negative residual variance was found in this model. The two cases were in Models 3 and 7 (for the WJ Reading Fluency and TOWRE Sight Word Reading, respectively). In both cases, the residual variance was restricted to zero and the model was re-fit. For both models, this resulted in an error-free convergence, and did not substantively shift the fit criteria. The fit results reported throughout this study reflect the adjusted model.

Table 1.5
Model Fit

Model	χ^2	df	CFI	RMSEA	SRMR	AIC
1. **Single factor**	655.5	27	.779	.193	.088	27875
2. **Bi-factor –silent word reading**	274.03	24	.912	.129	.052	27499
3. **Bi-factor – contextual word reading***	552.89	25	.814	.184	.086	27776
4. **Bi-factor – phonic decoding**	363.32	23	.880	.154	.062	27590
5. **Bi-factor – speededness**	232.42	22	.926	.124	.038	27462
6a. Bi-factor – battery	595.36	23	.799	.200	.084	27822
6b. Covariance – battery*	374.26	25	.877	.150	.073	27726
7. Bi-factor – mixed*	178.85	20	.944	.113	.035	27412

Note. Models 3 and 7 contain constraints restricting one residual variance to zero. Models 6b and 7 were fit post-hoc, after examining the previous models (explanation below).
χ2 = Chi square statistic, df = degrees of freedom, CFI = Comparative Fit Index, RMSEA = Root Mean Square Error of Approximation, SRMR = Standardized Root Mean Square Residual, AIC = Akaike Information Criterion

Bi-Factor Model – Phonic Decoding

The last trait factor analyzed separately in this model sequence was phonic decoding, which is an area of particular interest with this population, as phonic decoding has been shown to be an area of particular weakness. This model, like the other two trait models, did significantly improve fit from the single-factor (CFI=.880, RMSEA=.154, SRMR=.062), meaning the phonic factor was significant improvement in the sequence as well ($p<.01$), but of the three models in the trait sequence, the silent reading factor model was the best fit, and thus this factor was incorporated in a composite model with one of the method factors.

Bi-Factor Model – Speededness

The first of the two method factors modeled here was speededness, and using the same bi-factor structure as the trait models, we allowed the five speeded tests to load onto this factor as well as the overall reading factor simultaneously. The fit of this model was good - the best of the sequence thus far - and the speed factor accounted for quite a lot of the additional variance, resulting in improved model fit indices which met two of the three preferred cutoff points (CFI=.926, RMSEA=.124 SRMR=.038), with RMSEA for this and all models in the sequence being above the preferred cutoff point, though this measure alone can be sensitive to sample size and degrees of freedom.

Bi-Factor Model – Battery

To analyze the same structure with the other method factor - battery - we fit a model which allowed for the subtests which formally shared a test battery (the Letter-Word ID, Word Attack, Passage Comprehension, and Reading Fluency subtests of the WJ; the Phonemic Decoding and Sight Word Reading Efficiency subtests of the TOWRE), to load onto battery-specific factor in the case of the WJ tests, or to covary in the case of the two TOWRE tests, in order to create a method-focused bi-factor model based on battery. Thus, the two assessments which informally are from the same battery in that they are produced together and represent the same assessment method and format (the TOSCRF and TOSWRF) were initially not included in this model. The fit indices (CFI = .799; RMSEA=.200, SRMR .084) were not as good as the speededness model, and actually resulted in this being the worst fit of the trait and method bi-factor models tested thus far in the sequence. Importantly, the factor for the WJ subtests was not significant, and two of the four loadings for it were negative. However, the covariance between the TOWRE subtests was a significant source of identified variance. Thus, we fit a section version of this battery-focused model (see Model 6b in Table 1.5) which dropped the WJ factor, and instead added an additional covariance for the informal shared battery of the TOSCRF/TOSWRF tests. While the interpretation of this covariance is limited, as it is indistinguishable as a method factor from the silent trait they share, this updated version of the model does indicate that this additional pair does share a great deal of variance, whether that is attributable to a method effect or their shared assessed trait of silent fluency. Because this model was still a poor fit (CFI = .877; RMSEA=.150, SRMR .073), we can infer that the battery effect is not strong from either conceptual approach, both the paired tests and the larger shared battery.

Bi-Factor Model – Silent Word Reading and Speededness

Based on the results of models 2-6, seen in Table 1.5, a composite model was constructed with the two trait and method factors included from the five factors analyzed above which most improved the unidimensional model (see Figure 1.7). In selecting which factors to incorporate into the composite model, we proceeded with those which had the model fit indices at or closest to the preferred cutoff points of: CFI>.90; RMSEA<.05; SRMR<.08 (Barrett, 2007). The factors in the models which met this criteria were silent word reading (trait) and speededness (method), resulting in the composite model shown in Figure 1.7. The fit of this model was the best of the sequence (CFI = .944; RMSEA = .113; SRMR = .035), with two of three indices within preferred recommendations. This model suggests that among all of these reading-related outcomes, it is important to account not only for a general factor of reading, but also for speededness (several tests are timed) as well as for silent reading.

Figure 1.7: Method/Trait Factors

Discussion

The goal of this study was to examine the relations between reading skills in this population, by modelling traits and methods separately and in conjunction with each other using CFA models specified for an MTMM approach, which allowed us to more thoroughly describe the sources of variance in overall reading ability. The initial single-factor model, as well as the initial set of five simple bi-factor models, did not produce ideal fit results; however, the sequence did provide information about the potential factor structure which may explain variance across these assessments. Previous research indicated that the correlation across skills for this unique group of learners is low (Greenberg et al., 2011; Mellard & Fall, 2012; Sabatini et al., 2010), but questions about the relation across skills still remained, and the results of this study's factor models can begin to provide further information. Our sequence of models began with a single-factor model and a series of bi-factor models with one trait and method at a time, which indicated that the trait factor of silent word reading and speededness were the most important contributors to modeling overall word reading in this population. The outcome of this sequence was used to inform two post hoc models: an adaptation of the battery model, and a final composite model which included the trait and method which were the most important sources of variance. In some of the other factors assessed, interesting patterns were found.

For example, while phonic decoding was an area of weakness for struggling adult readers in this study as expected, in terms of the structure of reading skills analyzed in this sequence of factor models, phonic decoding was one of the least significant factors to add to the initial single-factor model, second only to the contextual-factor model. This means that while phonic decoding is not integrated with the other reading skills in this population, it also is not loading as one strong factor within the phonic decoding tests in this set of assessments, and they collectively do not account for much shared variance. On a practical level, this means these tests may not all be measuring one shared factor only. We know from the other models that the TOWRE Phonemic Decoding subtest and the TOSWRF, for example, may also be measuring a strong speededness factor.

Relations between the assessments indicated that silent word reading and speededness were the most important the trait and method factors - respectively - contributing to model fit, leading to their inclusion in the final composite model, which was the best fit in the sequence. These factors were significant in modeling overall reading performance and predicting outcomes on the battery of assessments used in this study. In terms of the sample studied here, we can infer that reading words silently and quickly – both being important components of fluency, are distinct skills within overall reading ability for this group of learners. Thus, ability in these two areas can account for significant variance in performance on the associated assessments outside of overall reading ability.

Within this sequence of models, the MTMM approach to interpretation allowed us to structure a concise and interpretable sequence of models which could offer insight on these assessments as well as the population of interest here. While the initial single-factor model results made it clear that these skills cannot be all interpreted as just one overall factor of reading, the series of bi-factor models which followed indicated that not all trait and method factors equally improved the unidimensional model. Thus, what we can learn from the factor models which did not result in an ideal overall model fit – such as the trait factor of contextual word reading or the method factor of battery of origin – is that these factors in conjunction with overall word reading do not fully explain the relation of assessed skills in this population. This is in keeping with prior research on measuring reading skills of struggling adult readers, and the model fit difficulties found in those studies (MacArthur, et al., 2010; Nanda et al., 2010). While the final model in the sequence which accounted for the most important trait and method factors to explaining the structure of the relation of skills assessed here was the best fit in the sequence, there is still room for improvement.

Limitations

While the MTMM approach to structuring and interpreting these models allows us to strategically analyze specific types of factors within and across these assessments, a limitation of this approach is that it does require a somewhat narrow definition of certain traits in order to categorize the tests. While the methods analyzed here (battery and speededness) are straightforward to categorize, some of the traits – specifically, silent reading and phonic decoding – can be a bit more complex than the test manuals' categorization, as tests may in reality, be measuring multiple overlapping traits. With the designation "silent" as a trait factor, there are tests such as the WJ Passage Comprehension, where completing a given item requires silent reading and an oral response. While we categorized this subtest as non-silent, because the scored response is oral, silent reading is still required to complete the item, and this potential multi-dimensionality is not captured in these models. Likewise, in categorizing assessments as measures of phonic decoding, the distinction can be unclear, as both phonic decoding and orthographic skills can be used to wholly or partially identify a real word or non-word.

A specific limitation is present in the two approaches to fitting the battery models. In the initial battery model, the TOSWRF and TOSCRF tests– while also informally from the same battery – were not set to covary. This was not only because the two tests are not formally produced as subtests under one battery, but also because that factor would then be indistinguishable from the factor of silent reading and difficult to interpret as a distinct method-based covariance (see Table 1.2). The overall conceptual structure of a battery-specific model includes the relation between these two subtests, but due to the complexity of the multiple relations across this battery of tests, not all cases could be simultaneously modeled in such a way to support a valid interpretation. In the adapted version of the model, once the WJ trait was found to be insignificant, this covariance was added, and was found to be a significant source of variance, but the limitation described above is a remaining concern, calling into question how the silent reading factor should be interpreted, as it may be accounting for specific trait variance, but it may also be accounting for a TOSWRF and TOSCRF factor based on their unique and shared test structure.

Future Research

Future research is needed to more closely examine the item-level structure of these tests in this population. Given the results of this study and prior studies indicating that on the test level, models of the relation of reading skills may not fit this population as expected with children reading at a typical age and grade level alignment, precedent is set that item-level models validating assessment design – such as the premise of unidimensionality – may also not hold with this population and these assessments. These questions must also be considered in the larger discussion of how these assumptions of test design are tied to the validity of the scoring process and interpretation. If, in the example of unidimensionality, these assumptions of test design do not hold, it has implications on what kind of models – like IRT – can be used to evaluate the measures. Future research is also needed to evaluate structural differences across groups within this population, such as native speaker status, age, and other demographic variables represented in this diverse group.

Implications

This study replicates and extends previous research validating the need for analysis of the structure and relations within and across assessments for the struggling adult reader population (Fritz, 2015; Greenberg et al., 2011; Greenberg et al., 1997; Mellard & Fall, 2012; Mellard et al., 2010, Nanda et al., 2010; Sabatini et al., 2010). The findings from this study are relevant to researchers of struggling adult readers, and can inform test selection decisions and test performance interpretations. Additionally, the relations across tests and skills modeled within this study allow for more accurate interpretation and discussion of scores for this population, and to better understand the role of constructs and methods in accounting for variance in outcome. Specifically, we see that some tests do not generalize to an overall reading factor, nor to a collective trait-specific factor (e.g. phonic decoding). For these tests, we should use caution when generalizing interpretation. Additionally, this research expands on trends of low correlation of skills seen in research with this population, to also demonstrate that this pattern may vary across the traits and methods of measurement, though still consistently lower than correlations of skills found with child readers. We do see strong shared factors for silent reading and speededness, separate from the overall reading factor, which indicates that there are complexities and variation in the relation of skills for this population, and that therefore these traits and methods may need to be modeled and interpreted as separate in research and practice with this population. The methodology from this study can inform future testing research for other atypical populations, or other content areas for adult learners.

Conclusions

The analysis of the structure of this study's selection of reading tests provides information about the relation between the tests themselves, and also about the performance of adult learners in these skill areas. Specifically, when using these assessments with this population, generalization of the results should be used with caution. Implications of what is found about the relation between traits which these assessments measure and the effect of method on the outcome and performance of this population offers insight on how these tests can be more effectively and accurately used and interpreted with an atypical population of learners such as struggling adult readers. Importantly, the results of this study serve as a caution to generalizing the results of a single measure of one reading skill to overall reading ability, for this population.

In conclusion, the results of this study extend the discussion of how reading tests assess a variety of skills in a sample of struggling adult readers, and how those skills relate to each other. The information about the intersection of this population and these assessments is examined within the context of a MTMM approach, which - as used here -outlines a method for examining the way both the skills being assessed and the method of assessment can be modeled for this group. The structure and sequence of these models allowed us to analyze and discuss the traits and methods being measured by these assessments separately. While the final model from this sequence included factors for speededness and silent word reading, which fit reasonably well, though still not an ideal fit, an important part of the information gained from this analysis is what did not fit well: the unidimensional model, and trait or method factors added alone to the unidimensional model. The outcomes of this study can inform the use of these tests, and the questions which arise from these results can inform future research on the relation of these skills and assessments for atypical populations.

References

Barrett, P. (2007). Structural equation modelling: adjudging model fit. *Personality and Individual Differences, 42*, 815–824.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Eason, S. H., Sabatini, J. P., Goldberg, L. F., Bruce, K. M., Cutting, L. E. (2013). Examining the relationship between word reading efficiency and oral reading rate in predicting comprehension among different types of readers. *Scientific Studies of Reading*, *17*, 199-223.

Eid, M., & Nussbeck, F. W. (2009). The multitrait-multimethod matrix at 50! *Methodology*, *6*, 71 doi: 10.1027/1614-2241.5.3.71

Fritz, C. M. (2015). *Modeling reading constructs with struggling readers at different ages.* (Ph.D. dissertation), Georgia State University, Atlanta, GA. Retrieved from https://scholarworks.gsu.edu/psych_diss/139

Greenberg, D., Ehri, L. C., & Perin, D. (1997). Are word-reading processes the same or different in adult literacy students and third–fifth graders matched for reading level? *Journal of Educational Psychology, 89*(2), 262-275.

Greenberg, D., Pae, H. K., Morris, R. D., Calhoon, M. B., & Nanda, A. O. (2009). Measuring adult literacy students' reading skills using the Gray Oral Reading Test. *Annals of Dyslexia*, *59*(2), 133–149. http://doi.org/10.1007/s11881-009-0027-8

Greenberg, D., Wise, J., Morris, R., Fredrick, L., Rodrigo, V., Nanda, A., & Pae, H. (2011). A randomized control study of instructional approaches for struggling adult readers. *Journal of Research on Educational Effectiveness*, *4,* 101-117.

Hammill, D., Weiderholt, J., & Allen, E. (2006). *Test of Silent Contextual Reading Fluency*. Austin, TX: PRO-ED.

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

MacArthur, C. A., Konold, T. R., Glutting, J. J., & Alamprese, J. A. (2010). Reading component skills of learners in adult basic education. *Journal of Learning Disabilities*, *43*(2), 108–121. https://doi.org/10.1177/0022219409359342

Mather, N., Hammill, D., Allen, E., & Roberts, R. (2004). *Test of Silent Word Reading Fluency*. Austin, TX: PRO-ED.

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology*, 4. doi:10.3389/fpsyg.2013.00169

Mellard, D. F., & Fall, E. (2012). Component model of reading comprehension for adult education participants. *Learning Disability Quarterly*, *35*(1), 10–23. https://doi.org/10.1177/0731948711429197

Mellard, D. F., Fall, E., & Woods, K. L. (2010). A path analysis of reading comprehension for adults with low literacy. *Journal of Learning Disabilities*, *43*(2), 154–165. https://doi.org/10.1177/0022219409359345

Mellard, D. F., Woods, K., & Fall, E. (2011). Assessment and instruction of oral reading fluency among adults with low literacy. *Adult Basic Education and Literacy Journal, 5*(1), 3-14.

Mellard, D. F., Woods, K. L., Deana Md, Z., & Alexandra Vuyk, M. (2015). Underlying reading-related skills and abilities among adult learners. *Journal of Learning Disabilities*, *48*(3), 310–322. https://doi.org/10.1177/0022219413500813

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.

Miller, B., McCardle, P., & Hernandez, R. (2010). Advances and remaining challenges in adult literacy research. *Journal of Learning Disabilities*, *43*(2), 101–107. https://doi.org/10.1177/0022219409359341

Nanda, A. O., Greenberg, D., & Morris, R. (2010). Modeling child-based theoretical reading constructs with struggling adult readers. *Journal of Learning Disabilities, 43*, 139-153.

Nanda, A., Greenberg, D. & Morris, R. (2014). Reliability and validity of the CTOPP Elision and Blending Words subtests for struggling adult readers. *Reading & Writing*, *27*(9), 1603–1618.

Nightingale, E., Greenberg, D., Branum-Martin, L. (2016). Selecting Fluency Assessments for Adult Learners. *Journal of Research and Practice for Adult Literacy, Secondary, and Basic Education, 5,* 18-29.

Pae, H., Greenberg, D., & Williams, R. (2012). An analysis of differential response patterns on the Peabody Picture Vocabulary Test-IIIB in struggling adult readers and third-grade children. *Reading & Writing*, *25*(6), 1239–1258.

Pae, H. (2012). Construct validity of the Pearson Test of English Academic: A multitrait-multimethod approach. *Educational Research and Evaluation*, *18*(6), 571-596.

Pearson. (2011). *Pearson Test of English Academic: Automated scoring.* Retrieved from www. pearsonpte.com/research/Documents/PTEA_Automated_Scoring.pdf

Platt, J. R., (1964). Strong Inference. *Science*. 146(3642), 347-353. DOI:10.1126/science.146.3642.347

Reynolds, C., & Kamphaus, R. W. (2007). *Test of Irregular Word Reading Efficiency*. Lutz, FL: Psychological Assessment Resources, Inc.

Sabatini, J., Sawaki, Y., Shore, J., & Scarborough, H. (2010). Relationships among reading skills of adults with low literacy. *Journal of Learning Disabilities*, *43*, 122-138.

Sechrest, L., Davis, M. F., Stickle, T. R., and McKnight, P. E. (2000). Understanding 'method' variance. *Research Design: Donald Campbell's Legacy*, ed. L. Bickman (Thousand Oaks, CA: Sage Publications), 63–87.

Tighe, E., and Schatschneider, C. (2014). Examining the relationships of component reading skills to reading comprehension in struggling adult readers: A meta-analysis. *Journal of Learning Disabilities*. Doi: 10.1177/0022219414555415 PMC 4532638

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2012). *Test of Word Reading Efficiency, 2nd Edition*. Austin, TX: PRO-ED.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive Test of Phonological Processing*. Austin, TX: Pro-Ed.

Wiederholt, J., & Bryant, B. (2001). *Gray Oral Reading Tests.* Austin, Texas: Pro-ed.

Woodcock, R., McGrew, K., & Mather, N. (2001). *Woodcock Johnson III Test of Cognitive Abilities*. Rolling Meadows, IL: The Riverside Publishing Company.

A MULTI-GROUP ITEM-LEVEL CONFIRMATORY FACTOR ANALYSIS OF WORD AND NON-WORD READING ASSESSMENTS ADMINISTERED TO STRUGGLING ADULT READERS

Review of the Literature

There are multiple skill components found within the general skill set we call reading, and one of the most foundational is word recognition, as it is a precursor skill to fluency and comprehension (Alvermann, Unrau, & Ruddel, 2013; Ehri, 2005). While many skills are important to the process of identifying words – such as phonemic awareness, orthographic word recognition, semantic knowledge, morphological awareness – the two main components considered here are the ability to orthographically recognize real words by using sight word skills, as well as the ability to phonologically decode real words and non-words. Assessments of word reading ability may measure one or both of those components of word reading. This study reports an analysis of measurement invariance with tests of word and non-word reading administered to native and non-native English speaking struggling adult readers, with an examination of both item and test information.

Before assessment outcomes can be interpreted or used to inform instruction, establishing that the assessment in question measures one single construct to be interpreted is a first step in the technical evaluation process. In this study we use structural equation modeling to assess the unidimensionality across and within three word and non-word reading tests, and consider what item-level information can tell us about how consistently the items measure a single construct of word reading across native and non-native speaking groups, as well as how phonic decoding and orthographic skills may contribute to or explain the patterns and relations found.

This study focuses on struggling adult readers. The latest survey of adult literacy skills sheds a harsh light on the reality of the number of adults in the United States who struggle with reading: nearly one in six adults score at or below the lowest level of reading proficiency (OECD, 2013). Research in the area of adult learners - specifically struggling adult readers - is quite sparse (NRC; 2012). The National Research Council's (2012) review mentions the need to understand more about the skill profiles and needs of struggling adult readers, as well as to better understand assessments for and with this population. Research on the word-reading skills of native and non-native English speaking struggling adult readers suggests that the relations between tests and the skills they intend to measure are not clear and consistent. One study successfully fit word reading within a five-factor reading skill model, alongside fluency, comprehension, spelling, and phonic decoding (MacArthur, Konold, Glutting, & Alamprese, 2010). However, other studies with this population have found poor model fit and low correlations between measures and constructs (e.g., Nanda, Greenberg, & Morris, 2010; Mellard, Fall, & Woods, 2010). The question of what construct or constructs these tests measure is part of the larger discussion of how word and non-word reading components relate to reading skills overall for this population.

Before describing the current study, a brief literature review will be provided to discuss the importance of word reading, and the contributions of phonic decoding and orthographic skills to word reading. Literature on struggling adult readers will be highlighted. Adults with reading difficulties include individuals who are not native speakers of English, and therefore a brief review of word reading and adult non-native speakers will also be addressed.

Importance of Word Reading

Word identification is an important foundational component of the general process of reading skill acquisition (Ehri, 2005). For example, McIntosh and colleagues (2006) found that with typically developing first-grade children, lower word-reading abilities early on were correlated with other literacy struggles including reading comprehension later in their reading development. Other studies with both elementary and secondary students have also demonstrated the relation between word reading skills and overall reading comprehension and fluency abilities (e.g., Paige, Rasinski, Magpuri-Lavell, & Smith, 2014; Cummings, Dewey, Latimer, & Good, 2011). In typically developing learners, the link between word recognition and reading comprehension may be attributed to the reduction in cognitive demand offered by skilled word recognition (Rapp, van den Broek, McMaster, Kendeou, & Espin, 2007). Rapp and colleagues explain that as word recognition skills move beyond active identification to automatic identification, the skill of overall reading fluency develops, as the cognitive demand of individually recognizing and "sounding out" each word declines. Thus, as learners develop the ability to recognize words quickly by sight, the cognitive demand of word recognition lowers, and the effort in reading can be concentrated on comprehension, rather than on phonological decoding of words.

In the struggling adult reader population, word identification is correlated with fluency and with reading comprehension (e.g., Sabatini, Sawaki, Shore, & Scarborough; 2010; Eason, Sabatini, Goldberg, Bruce, & Cutting, 2013). Sabatini et al. (2010) recommend that any screening or assessment of adult learners for research or educational purposes should be heavily weighted on word recognition skills, as their research indicates that word recognition skills are predictive of overall reading, because they are precursor skills to reading fluency. In Sabatini's study, word recognition skills were measured by both orthographic and phonic decoding measures, and within their model were found to be equally as predictive of passage comprehension as a language comprehension factor.

Word Reading Approaches

Measuring word reading out of context (words presented in isolation, rather than in a sentence) provides a purer measure of word identification ability, than measuring word reading in context (sentences/passages) in that it reduces external variance from the aid of the sentence/passage context. There are multiple methods to read real words out of context. If a real word is familiar, typically it is recognized automatically by sight, which we refer to in this study as sight word reading or orthographic word reading. If a word is not automatically recognized by sight, one can decode parts of the word by using phonic decoding (in some studies referenced here this can also be referred to as phonological decoding), or by using an analogical approach where parts of a word are recognized by establishing a mental analogy to familiar words with shared parts. Guessing is also another strategy for identifying either real words or non-words, and can be used in full isolation, or more likely in combination with some partial approach to identification via phonic decoding or orthographic awareness. In this study, analogical word reading skills and guessing are not examined nor modeled separately from orthographic or phonic decoding.

The terminology used for reading real words and non-words can vary, and the distinction between which of the skills is necessary and used to read a given word or non-word is difficult to make, since real words can be read using awareness of phonemes and whole word recognition. Kilpatrick (2015) outlines a method of conceptualizing the difference between the approaches to word reading, explaining that real words can be processed through phonic decoding (sometimes referred to as phonological word reading) or word recognition (sometimes referred to as orthographic word reading). Based upon this method of conceptualizing the skill sets of reading words using awareness and mapping of phonemes to sounds and reading words using whole word recognition using memory and sight, this study will primarily use the terms phonic decoding and orthographic word recognition for the conceptual discussion of these skills, and word and non-word reading for the practical application of categorizing and modeling assessment items.

Phonic Decoding and Orthographic Word Recognition

The two categories of word reading discussed in this study are phonic decoding and orthographic word recognition. The research on phonic decoding is clear that this skill is critical to multiple areas of reading ability (Bradley & Bryant, 1983; Mann, 1993; Wagner, Torgesen, & Rashotte, 1994). Phonic decoding plays an important role not only in allowing the reader to read words, but also in the post-identification process of memory and comprehension (Leinenger, 2014). Myers and Robertson (2015) also demonstrate that phonic decoding is predictive of individual word reading skills, in addition to being one of the strongest predictors of sentence comprehension skills. Orthographic word recognition can be defined as the ability to accurately identify words by sight, a process involving memory and recognition (Ehri, 2014). Research has shown that orthographic word recognition skills are an important predictor of both fluency and comprehension (Alvermann et al., 2013). Orthographic word recognition involves identifying each word as a unit by using sight memory to connect the string of letters with the sounds and meaning associated with the word (Ehri, 2014). The process of developing orthographic word recognition skill, in both children and adults, is linked to exposure (Bosse, Chaves, Largy, & Valdois, 2015, Grainger & Hannagan, 2014). Adults who struggle with literacy tend to perform better on orthographic measures as compared to their performance on phonological measures (e.g., Sabatini, et al; 2010).

Much research on phonic decoding has indicated that it is an area of particular weakness for struggling adult readers (e.g., Greenberg, Ehri, & Perin, 1997; Greenberg, Ehri, & Perin, 2002; Greenberg et al., 2009; Sabatini, et al; 2010). Greenberg et al. (1997) found that for adult struggling readers, unlike the patterns seen in reading-matched children, phonological decoding was a much greater struggle than orthographic decoding for their adult learners. Greenberg et al, (2002) found adults with low reading skills opted for orthographic and visual decoding strategies rather than phonological decoding more often than children matched by reading level. Likewise, Thompkins and Binder (2003) found that compared to typically developing children, struggling adult learners relied more heavily on orthographic skills than phonological decoding. Binder and Borecki (2008) looked at how skilled adult readers and struggling adult readers performed on a task of identifying incorrect homophone usage (participants identified when words with the same pronunciation but different spelling/meaning were used incorrectly). They found that skilled adult readers performed similarly on orthographic and phonological versions of the task, whereas the struggling adult readers were far slower at identifying the incorrect word under the phonological condition. As a final example, Tighe and Schatschneider (2016) noted in a meta-analysis of multiple studies with this population that real word reading was repeatedly observed with a stronger correlation to reading comprehension as compared to the correlation between non-word reading and reading comprehension.

Researchers have also explored whether struggling adult readers' reading related skills form a well-integrated reading system. Mellard and Fall (2012) found in their sample of adult learners who struggle with reading that their skills sets of phonological awareness/decoding, sight word reading, and comprehension were correlated, but were not as well integrated as we would expect in typical readers. In addition, similar to the research already described, a principal components analysis revealed that struggling readers relied more heavily on word reading skills and memory, while more advanced readers drew more evenly from multiple skills (Mellard & Fall, 2012). Other studies with struggling adult learners show a lower correlation between performance on phonological and orthographic decoding assessments than with children (Greenberg et al., 2011; Joseph, Torgesen, Wagner, & Rashotte, 1999). Nanda, Greenberg, and Morris (2014) similarly found that models of reading skill correlations in children do not fit as well for adults. Overall, these findings suggest that, across literacy skills, tasks among struggling adults are not as highly related as we might expect, suggesting a literacy system which is not well integrated.

Non-native English Speakers and Word Reading

Within the population of adult learners, there are both native and non-native English speakers, and some research is available on the potential complexities this may add to the non-homogeneous adult learner group. In terms of overall ability, research comparing native and non-native speakers of English in this population of struggling adult readers has found that while non-native speakers outperform native speakers on measures of decoding and phonological awareness, they perform lower in the area of language comprehension (MacArthur et al, 2010). Additionally, Davidson & Strucker (2002) found that non-native English speakers more closely resembled normally developing readers in terms of patterns of correlations between developing reading skills, whereas the native English speakers more closely resembled children with learning disabilities. While there is little advanced item-level research with struggling adult readers examining the effect of native speaker status there is certainly cause to question whether native speaker status may play a role in how word reading tests/items function with these learners.

Assessing Struggling Adult Readers

Interpretation of the research described above relies on the assumption that the assessments that were administered in the studies are psychometrically appropriate for struggling adult readers. However, when evaluating the technical quality and validity of the outcomes from assessments administered to this population, there are key assumptions to re-check in the context of this population, the most important of which are the assumptions of unidimensionality and intended item difficulty ordering/structure. Although this study involves an item-level analysis, for all three assessments used here, one single outcome score is reported in practice, which indicates the underlying assumption that one construct is measured by the assessment. Additionally, in each assessment, specific basal and ceiling rules are applied in administration (see further information in the Measures section) which rely on the assumption of increasing item difficulty to hold for this population in order for the scores based on the basal and ceiling rules to be a valid outcome with limited error based on the imputed correct/incorrect scores at the extremes of the measure. As overarching guidance for these issues, the *Standards for Educational and Psychological Testing* states: “Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population” (AERA, 2014, p. 195). In examining the factor structure of tests within and across groups, this study begins to answer questions tied to the validity of the use and interpretation of assessments for this population.

As noted above, studies with struggling adult readers have demonstrated that the correlation of reading skills such as word reading and comprehension was lower in struggling adult readers as compared to typically developing children; but the implications of these correlations can be unclear, without certainty that certain assessment features (unidimensionality, increasing item difficulty, etc.) function as intended with this population (Greenberg et al, 2011; Greenberg et al, 1997; Mellard et al, 2010; Nanda et al, 2010; Sabatini et al, 2010). In addition, reading assessments have not been specifically normed with this population of adults who struggle with reading, and researchers have found difficulties with some of the reading tests and this specific population. For example, in a study in which the Gray Oral Reading Test (Wiederholt & Bryant, 2001) was administered to adult struggling readers, the results indicated that this assessment did not function with adult learners the way it did when normed on children (Greenberg, Pae, Morris, Calhoon & Nanda, 2009). This is similar to the findings conducted by Nanda and colleagues (2014) on the Comprehensive Test of Phonological Processing (CTOPP, Wagner, Torgesen, & Rashotte, 1999); the validity and reliability of the measure were lower than what was reported in the manual with children. As another example, Pae and Greenberg (2012) conducted an item-level analysis on the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 2007) to determine if item bias or differential item functioning was present. They found for their sample of struggling adult readers, there were some examples of item misfit, which when removed, improved the unidimensionality of the test. The results in the studies mentioned above indicate there is cause to examine how reading – in the case of this study, word and non-word reading – tests function on this unique population.

Structural Equation Modeling (SEM)

Some previous studies (Tighe, Schatschneider, Crepaldi, & Tomás, 2016; Hannon, 2012; Babayiğit, 2015) with adult learners have used Structural Equation Modeling (SEM) to model the relations between reading tests and the abilities they may represent, because this methodology allows for an examination of the variance in outcomes through freely estimating or restricting the relation between possible factors accounting for variance. For example, Tighe and her colleagues (2016), using SEM, found a strong relation between identifying both real word decoding and reading comprehension in a group of struggling adult readers, but a moderate relation between both phonological awareness/pseudo-word reading and reading comprehension. Hannon (2012) used SEM with proficient adult readers and found lower-order processing like word-reading was not strongly correlated with higher order comprehension skills. These different findings indicate that word reading constructs may relate differently in proficient and non-proficient readers.

In addition to identifying the structural relations between skills and assessments, SEM can also be used to model and compare any structural differences across groups of learners. In a specific example of this application of the method, a study which examined word reading in children found that native speakers performed better than non-native speakers in the area of oral word reading, but not in word reading overall (Babayiğit, 2015). This current study likewise used an SEM framework – specifically, a Confirmatory Factor Analysis (CFA) approach – to first evaluate the factor structure of these items, and then compare model fit and outcomes across native and non-native groups in order to identify any differences in item-level function.

Multitrait/Multimethod (MTMM)

When looking at the relation between items and constructs that tests measure, the Multi-Trait/Multi-Method (MTMM) approach is a helpful framework which separates the influences upon measures by traits and methods (Campbell & Fiske, 1959; Maul, 2013). “Trait” means the skill intended to be measured and “method” means the mode, test, or process by which the measure is administered. Assessments may measure the same trait using the same method, the same trait using a different method, a different trait using the same method, or a different trait using a different method. Alternatively, “methods” might also represent sub-traits or specific factors (e.g., speed) which are relevant to some tests but not others (Maul, 2013).

MTMM approaches are commonly used by researchers (Dickinson & Adelson, 2016; Shermis & Long, 2009; Nightingale, Greenberg, & Branum-Martin, 2016) to provide a framework for the interpretation and evaluation of analyses of or related to construct validity, and can be applied both to simple correlation methodologies and advanced structural models. For example, Dickinson and Adelson (2016) used an approach informed by MTMM theory to compare assessment outcomes’ prediction of college success. They found that the outcomes categorized as trait-based accounted for more of the variance than method-based outcomes. Similarly, with children, MTMM has been used to look at how item types (multiple choice/performance) and content areas (reading/writing) correlate to each other, and to determine if the variance is attributed to the difference in trait measured, or just differences in the method. For example, in a study reported by Shermis and Long (2009), their results indicated that shared methods (item type/format) of standardized subtests accounted for a larger portion of the variance in child performance than shared constructs between subtests. In a similar fashion, the current research study investigated results based on shared methods and shared traits decoding.

Only one previous study has used the MTMM approach with adult struggling readers. That study indicated that a MTMM approach reveals a consistent lack of convergent and discriminant validity for reading assessments – including word reading assessments measuring phonic decoding and sight word recognition (Nightingale, Greenberg, & Branum-Martin, 2016). Both in the method of assessment administration, as well as the trait measured, that study found a lower correlation of outcomes across measures as compared to the previously reported relations between these assessments when administered to typically developing children. While the focus of the present study does not directly consider the convergent and discriminant validity of assessments, the outcomes of the CFA models used in this study are likewise interpreted using the MTMM framework to isolate trait and method related variance.

Purpose

This study examined three word and non-word reading assessments administered to a sample of native and non-native English speaking struggling adult readers tested in the United States and Canada. Based on the literature discussed above, there are still many questions to be answered about how word and non-word reading assessments function with adult learners, both at the test level and the item level. It is also important due to the linguistic diversity of this population to examine how language status may influence item and test level performance. The need for applied advanced statistical models in this field is present, and the purpose of this study is to address one part of that by exploring the function of three word and non-word reading tests in this sample as a whole as well as by native speaker status. While research indicates that there is a pattern of higher orthographic relative strengths and lower phonic decoding skills in adult struggling readers (Greenberg et al, 1997, Sabatini et al; 2010), it is unclear how these patterns may present on an item level and also how they may apply to the different groups within the wide population of adult struggling readers, including native and non-native speakers. Specifically, it was hoped that this study would provide information on the literacy skills in this population, and also present an interesting look into the complex structure of word and non-word reading tests, and how those concepts are presented and measured in these assessments.

The three tests that were chosen for this study are the Test of Irregular Word Reading Efficiency (TIWRE; Reynolds & Kamphaus, 2007) and the Woodcock Johnson subtests of Letter-Word Identification (WJ-LWI) and Word Attack (WJ-WA; Woodcock, McGrew & Mather, 2007). The Woodcock Johnson subtests are typically used by adult literacy researchers and capture both phonological and orthographic skills (Mellard & Fall, 2012; Sabatini et al, 2010). The TIWRE measure is not as frequently used, but we included it in this study to consider if there is a method effect based on assessment battery, as it is not part of the Woodcock Johnson battery. Performance on the traits that these assessments measure, as well as the measurement method, is of interest here, and in addition to the examination of the whole group's performance on these items, we also look at how group differences like native speaker status may affect model fit and item indices.

Using a series of CFA models, we can confirm the applicability of a single-factor (unidimensional) model within and across these measures, and also look at how the relation between the assessments and the traits/constructs overlap. The models used in this study are constructed based on the MTMM approach to modeling and interpreting multiple factors, in order to discuss what model fit means in terms of what the tests are measuring. Outcomes of the models can then be examined across native and non-native speakers to determine if the structure found is equally appropriate for both groups.

Exploratory Research Questions:

Research Question 1: To what extent do the items of these three tests measure general word reading ability versus trait-specific and method-specific factors when administered to adult struggling readers?

Research Question 2: To what extent might the measurement structure be equivalent in both native and non-native speakers of English?

Methodology

Participants

Data were collected in a larger study (IES grant #R305C120001; PI: Daphne Greenberg) between 2012 and 2015 from 931 participants who attended multiple Adult Basic Education classes in two large urban cities in the southeast US and Ontario, Canada. Participants were recruited from classes targeting reading skills between the 3rd and 8th grade levels. The three tests analyzed in this study were part of a larger battery of assessments and surveys administered by trained graduate research assistants.

Sixty-seven percent of the sample were native speakers of English. Participants were asked a series of questions to determine their English speaking status. To qualify as a non-native speaker, participants had to self-identify as a fluent speaker of a language other than English, and then list a language that is not English as the first language they learned to speak. Participants who did not speak any language other than English fluently and learned English as their first language were classified as native English speakers. As can be seen from Table 2.1, the non-native English speakers had a very diverse range of first languages, with Spanish being the most common, and French coming in second, followed by Arabic, German, and Creole, with all of the remaining 40+ languages represented in our sample having five or fewer representative first-language speakers.

Table 2.1
First Language Spoken by Non-native English Speakers

Language	Frequency	Percent
Spanish	32	20%
French	13	8%
Arabic	11	7%
German	7	4%
Creole	6	4%
Amharic	5	3%
Edo	5	3%
Persian	5	3%
Punjabi	5	3%
Urdu	4	3%
Farsi	3	2%
Patois	3	2%
Somali	3	2%
Twi	3	2%
Other	55	34%

Note: All other languages represented in the sample had two or fewer first-language speakers.

Participants were between the ages of 16 and 73 (mean = 37, SD = 14). Sixty-seven percent were female, and 71% of the sample was unemployed. The participants were Black (64%), White (28%), Asian (6%), Native American/First Nation (2%) and other (<1%). Table 2.2 disaggregates the demographic information by native speaker status. An analysis of variance (ANOVA) indicated that the relation between age and native speaker status was insignificant ($F = .36$, $p = .55$). In addition, chi-square goodness of fit testing for significant difference across native speaker groups indicated no significant differences for gender ($\chi^2 = 1.26$, $p = .26$) and employment status ($\chi^2 = 6.29$, p = .18). However, a chi-square goodness-of-fit test for race was found to be significant ($\chi^2 = 128.19$, p < .01).

Table 2.2
Demographics by Native Speaker Status

		Native	Non-native
Age	Mean Age	39	38
	SD	15.1	11.8
	Age Range	16-73	18-66
Gender	Male	34%	31%
	Female	66%	69%
Race	Black	73%	44%
	White	22%	32%
	Asian	2%	23%
	Native American/First Nation	2%	1%
	Other	<1%	<1%
Employment	Employed	24%	40%

Measures

The *Woodcock Johnson III Test of Achievement* Letter-Word Identification subtest (WJ III-LWI; Woodcock et al., 2007) is a measure of oral word identification, comprised of words which can be read by using phonic and/or orthographic skills. The test was normed on individuals aged 2 through 99 years old. Examinees are presented with lists of words which they are asked to orally identify, and this continues through the end of the test of seventy-six items or until the examinee incorrectly identifies six consecutive words.

The *Woodcock Johnson III Test of Achievement* Word Attack subtest (WJ III-WA; Woodcock et al., 2007) is a measure of phonic decoding, comprised of phonological blends that are not real words, but must be orally identified by the examinee using phonic decoding rules. Like the WJ Letter-Word Identification subtest, the norming population for this test included a range of ages 2 through 99 years old. Examinees are presented with lists of non-words which they are asked to orally identify, and the test concludes when the examinee incorrectly identifies six consecutive non-words in a list or reaches the final item of the thirty-two non-words.

The *Test of Irregular Word Reading Efficiency* (TIWRE; Reynolds & Kamphaus, 2007) is a measure of sight word reading. The examinee is presented a list of items, which they are asked to orally identify until the examinee completes the list or incorrectly identifies four consecutive items from the list of fifty items (words). The test has been normed on individuals aged 3 to 94 years old.

All three tests are orally administered and require oral responses. Eleven items from each measure were included in our models, after trimming the items with limited variability in responses. Specifically, items with less than 10% variation in responses were trimmed, and then for each measure, the sequence of 11 consecutive items with the most variation in responses were selected. WJ Letter Word ID items 40-50, WJ Word Attack items 15-25, and TIWRE items 30-40 were included in this study.

In this study, the WJ Letter-Word ID and the TIWRE measures are conceptualized as relying heavily on orthographic skills, and the WJ Word Attack as relying heavily on phonic decoding skills. Because completion of items on all three tests can rely on both orthographic word recognition and phonic decoding, the primary designation for the purpose of the trait measured in these models is based on the types of items presented in each test: real words, or non-words. The WJ Letter-Word ID presents words that can be recognized by sight and/or identified with the aid of phonic decoding. The TIWRE includes phonetically-irregular words that primarily rely on sight word reading skills, although phonic decoding can help with some aspects of reading the words. The WJ Word Attack is perhaps the "cleanest" of the tests included in this study and primarily relies on phonic decoding skills. Since all three tests include items which can be correctly identified using some combination of either/both orthographic and phonological skills, for the purposes of trait modeling, tests were classified as either measuring real words or non-words.

Analytical Approach

Research Question 1: To what extent do the items of these three tests measure general word reading ability versus trait-specific and method factors when administered to adult struggling readers?

In order to look at the relation of the items as well as the relation between the three tests while taking into consideration trait and method, we used Structural Equation Modeling (SEM) to model the hypothesized relation between tests/items which measure the same trait. To answer this question we assessed the fit of single-factor item-level within-test models for each of the three word reading tests/subtests, followed by an across-test single-factor model to determine whether all three tests measured one single construct as well. Model fit criteria used within this study include the Comparative Fit Index (CFI > .95), Tucker-Lewis Index (TLI > .95), Root Mean Square Error of Approximation (RMSEA < .06), and Standardized Root Mean Square Residual (SRMR < .09), and are evaluated based on Hu and Bentler (1999) criteria for evaluating model fit. Data were processed and coded using SAS 9.4. SEM was fit using Mplus (Muthén & Muthén, 2018). In addition to evaluating model fit and the strength of the item loadings, to evaluate the within-test structure, item thresholds, which are related to the difficulty of the item, are also evaluated to assess the intended test structure of increasing item difficulty. We then fit a bi-factor model informed by the Schmid and Leiman bi-factor solution (1957) to measure how these tests/items fit within the overall word-reading construct as well as the constructs of phonic decoding and orthographic skills. Using this model with item level data allowed us to take into consideration conditional dependence aside from which test the item belongs to, such as in the case where one test may have both words which can be decoded with both phonic and orthographic skills (Gibbons et al., 2007). Additionally, a similar structure was used to look at the effect of method (test battery) on item performance. This can be assessed at the same time as trait, which allowed us to compare the effect with the method simultaneously.

Research Question 2: To what extent might the measurement structure be equivalent in both native and non-native speakers of English?

The models fit for the first research question were then fit within a multi-group CFA by native speaker status to look for any differences in model fit based on group status. To address the question of whether this structure was equivalent across groups, we fit a sequence of measurement invariance models (baseline/configural, metric, and scalar) in order to test for item bias versus genuine group differences for native versus non-native speakers of English. The item responses were analyzed using multiple group CFA, to isolate item-level measurement differences versus person-level differences due to group membership (i.e., language background). While the configural model, as the baseline, estimated free factor loadings across groups, in the metric model we restricted the factor loadings to be equal, to isolate any group-specific invariance. Building on this, we then also restricted the item intercepts to be equal across groups in the scalar model. Testing measurement equivalence in this way can reveal practical considerations for testing native and non-native English-speaking struggling adult speakers (e.g., some items are inordinately difficult), or might even suggest differences in cognitive processing or learning (e.g., the factors have differing means or structures).

Results

Before examining the measurement properties of the three assessments used in this study, we considered the overall performance of our sample (see Tables 2.3a - 2.3c). Both groups performed the lowest on the non-word reading test, and highest on the irregular word reading test. When comparing native and non-native readers, we found that the non-native group performed better than the native group on non-word reading, but performed lower on the measures with real words. Differences in performance by native speaker status were tested for significance using an ANOVA model, and significant differences were found only for the WJ Word Attack ($F = 4.61$, $p = .03$). For both the TIWRE ($F = .49$, $p = .48$) and WJ Letter-Word Identification ($F = 1.49$, $p = .22$), although a mean shift was observed, the differences were insignificant across groups. The tables below summarize this performance within and across groups, using the raw mean, standard score mean, grade equivalent mean, and also noting the maximum possible points per test.

Table 2.3a
Performance on Reading Assessments – Total Sample

Tests	n	Raw mean	Max possible	SD	Standard Score Mean	GE mean
Test of Irregular Word Reading Efficiency	519	37.83	43	5.23	-	6.7
WJIII: Letter-Word Identification	869	54.15	76	8.69	82.52	5.3
WJIII: Word Attack	891	16.02	32	7.45	80.99	3.1

Note: WJIII=Woodcock Johnson III Test of Achievement/Cognitive Abilities. Standard scores are not generated for the Test of Irregular Word Reading Efficiency

Table 2.3b
Performance on Reading Assessments – Native

Tests	n	Raw mean	SD	Standard Score Mean	GE mean
Test of Irregular Word Reading Efficiency	382	37.99	5.07	-	7.1
WJIII: Letter-Word Identification	543	54.04	8.81	81.99	5.3
WJIII: Word Attack	556	15.64	7.57	80.16	3.1

Note: WJIII=Woodcock Johnson III Test of Achievement/Cognitive Abilities. Standard scores are not generated for the Test of Irregular Word Reading Efficiency.

Table 2.3c
Performance on Reading Assessments – Non-native

Tests	n	Raw mean	SD	Standard Score Mean	GE mean
Test of Irregular Word Reading Efficiency	136	37.35	5.62	-	6.7
WJIII: Letter-Word Identification	276	54.49	7.62	83.57	5.3
WJIII: Word Attack	285	16.62	6.84	82.65	3.3

Note: WJIII=Woodcock Johnson III Test of Achievement/Cognitive Abilities. Standard scores are not generated for the Test of Irregular Word Reading Efficiency.

To answer our research questions, as explained in the Measures section, 11 items from each measure were included in our models, after trimming the items with limited variability in responses.

Research Question 1: To what extent do the items of these three tests measure general word reading ability versus trait-specific and method factors when administered to adult struggling readers?

In order to look at the relation of the items as well as the relation between the three tests, we put together a series of models using SEM – specifically, CFA models - using Mplus 8.1. To first address the unidimensionality question of whether all the items within these tests measure a single factor together, we ran one-factor CFA for each of the three tests, in which all items loaded onto one overall factor, after trimming the first and last few with less than 10% variance. The single-factor item-level models fit well separately for all three assessments (see Table 2.4, Figures 2.1-2.3). To further examine model fit, and to learn more about individual item performance and characteristics, we also considered the item response parameters of item difficulty. The results here indicated that for all three tests, there are a couple items outside of increasing order of difficulty - one in particular each for the orthographic/real-word-based assessments - WJ Letter Word ID (item 8) and TIWRE (item 4) - and a more gradual increase for the phonic decoding/non-word assessment - WJ Word Attack. However, there were not structural patterns deviating from the expected pattern of increase (see Figure 2.4). Looking back at the item loadings on the single-factor models, however, we see a few examples of loadings lower than expected (such as the first WJ Word Attack item loading, and items two through four of the WJ Letter-Word ID), but without an apparent relation to difficulty differences or anomalies (see Figure 2.1-2.3). From this we can infer that while the difficulty sequencing may not translate directly to struggling adult readers from the design for typically developing children, the few items slightly out of expected difficulty patterns do not practically influence the intended unidimensionality of the measures.

Table 2.4
Model Fit – Full-Group Single-Test Models

Models	CFI	TLI	RMSEA	χ^2	df	SRMR
WJ-Word Attack 1-Factor	.991	.988	.036	116.253	44	.036
WJ-Letter Word ID 1-Factor	.989	.986	.036	113.071	44	.043
TIWRE 1-Factor	.991	.989	.033	68.911	44	.050

TIWRE=Test of Irregular Word Reading Efficiency, WJ=Woodcock Johnson III Test of Achievement/Cognitive Abilities.

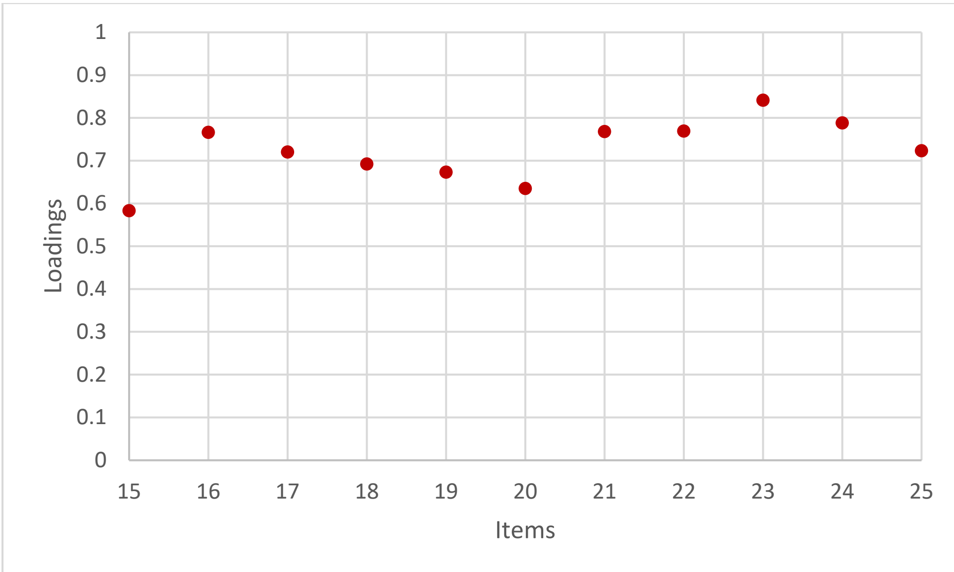

Figure 2.1. Standardized Item Loadings – WJ-Word Attack

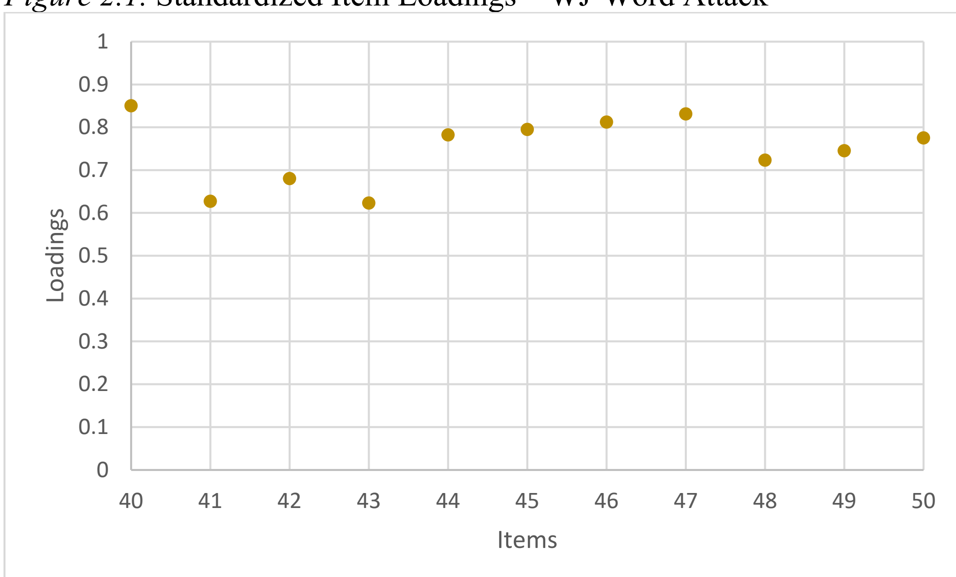

Figure 2.2. Standardized Item Loadings – WJ-Letter-Word ID

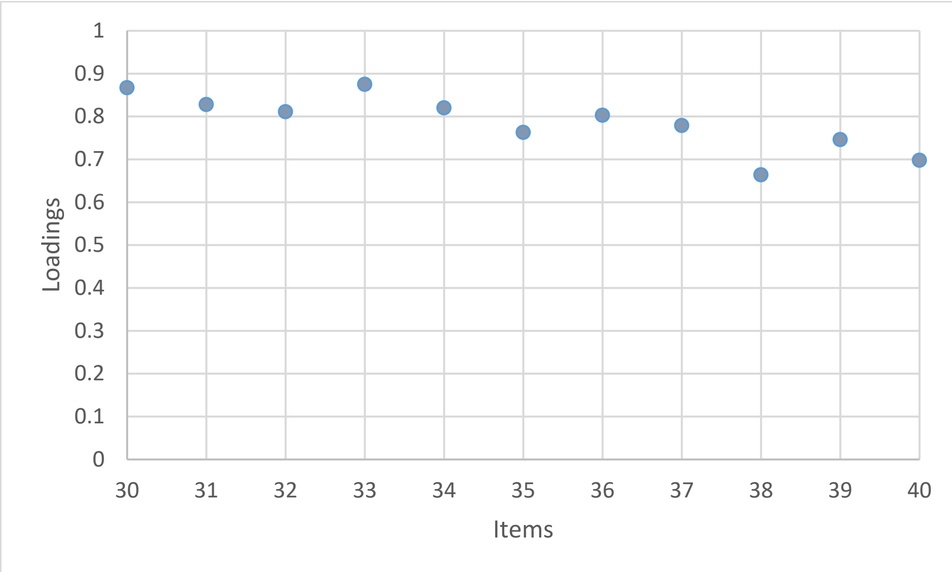

Figure 2.3. Standardized Item Loadings – TIWRE

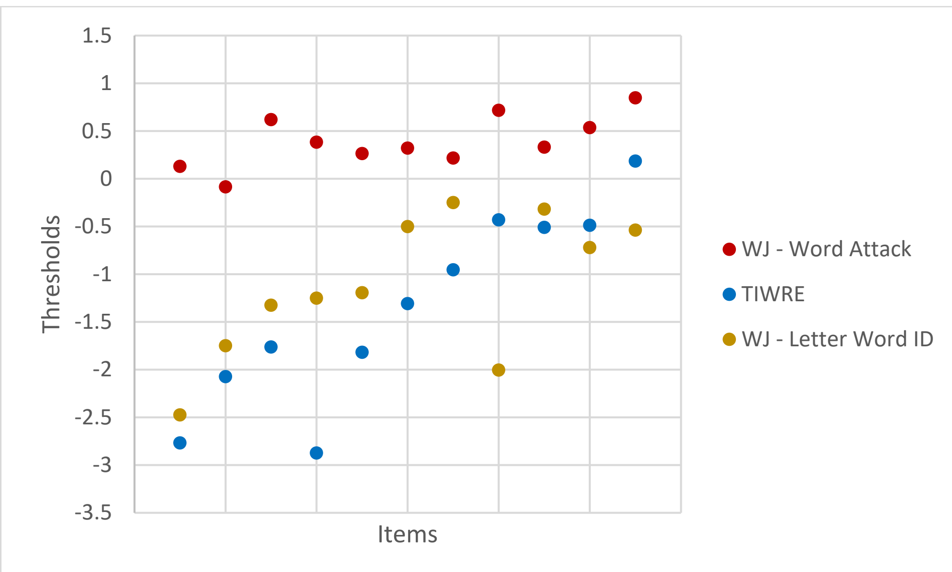

Figure 2.4. Unstandardized Item Thresholds

While the one-factor models for all three tests - both with real words (TIWRE and WJ Letter Word ID) and non-words (WJ Word Attack) - had a good model fit, the differences which were observed in model fit and loadings alone were not strong or informative, and thus prompted for further joint models examining the differences in decoding phonic or orthographic skill based items – that is: items of real words or non-words. First, to measure joint test structure and how well these three tests overall measure the same factor of word reading, all three tests jointly were fit to a full item-level one-factor model (see Figure 2.5). This was a not an ideal fit (see Table 2.5), and certainly not as good of a fit as the within-test models, indicating that the tests are assessing discrete factors between tests.

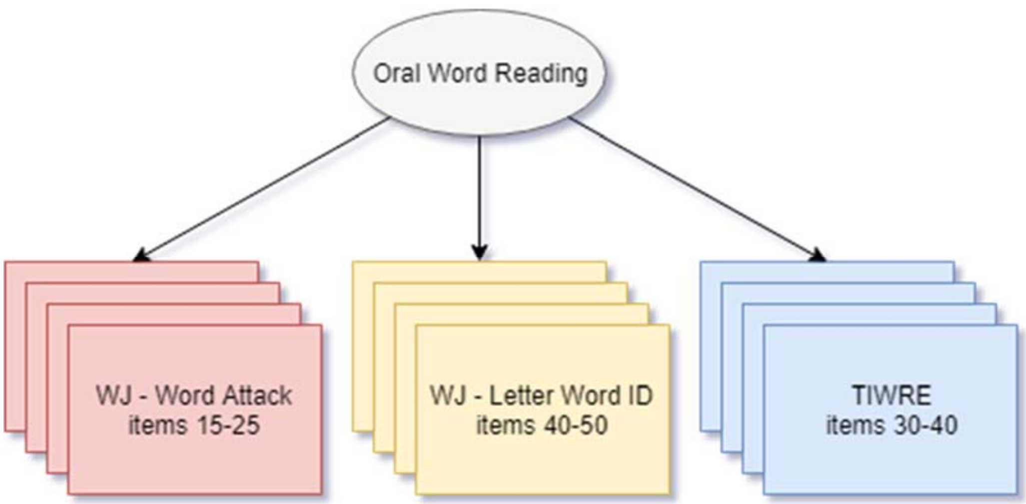

Figure 2.5. Single-Factor Model of Word Reading Across Tests

To further analyze the impact of the items' orthographic/phonic decoding status, we fit a bi-factor model to measure how these tests/items fit the overall word-reading construct when also accounting for whether an item is a real word or non-word (see Figure 2.6). In this model, accounting for whether an item is a real or non-word as a separate factor from overall word reading (see Table 2.4) improved fit from the joint one-factor model (CFI = .983; TLI = .981; RMSEA = .025; SRMR = .067). After completing this model structure based on trait (real word status), the same was done based on method (test battery), by fitting a model with a factor for the Woodcock Johnson battery, for the two subtests in this set which are from that battery (see Figure 2.7). This model was a good fit as well (CFI = .989; TLI = .987; RMSEA = .021; SRMR

= .058), close to the overall fit of the trait-specific bi-factor model (see Table 2.6), with still room for improvement on both.

It should be noted that the improvement in fit observed between the unidimensional model and the bi-factor models, while significant, does not meet Chen's (2007) criteria for model improvement, and the more parsimonious model could be retained. However, for the purpose of thorough investigation of the traits and methods represented in these tests, the bi-factor models will be retained for the second step in this analysis, in order to consider measurement invariance for both models.

Table 2.5

Model Fit – Full Group Across-Test Models

Models	CFI	TLI	RMSEA	χ^2	df	SRMR
Full 1-Factor	.976	.975	.029	1028.75	495	.076
Real Word Bi-factor	.983	.981	.025	857.829	475	.067
WJ Bi-factor	.989	.987	.021	698.273	442	.058

Note: WJ=Woodcock Johnson, *TIWRE=Test of Irregular Word Reading Efficiency*
The $\chi 2$ is calculated using the Satorra and Bentler mean-and-variance-adjusted (scaled) chi-square method (Satorra & Bentler, 2010).

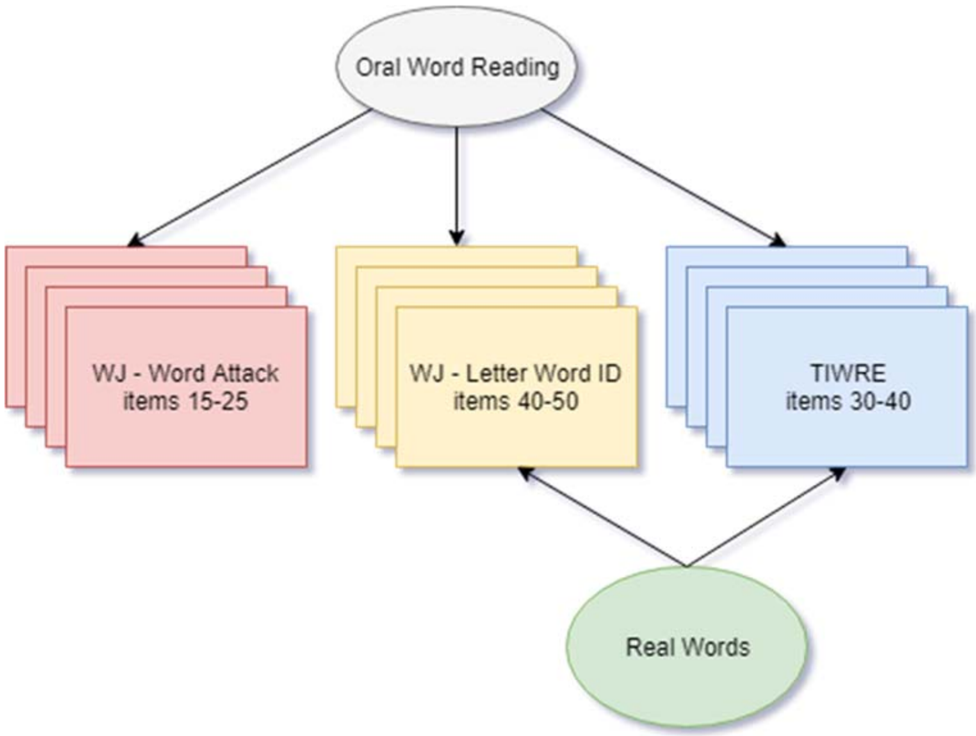

Figure 2.6. Bi-Factor Model with Trait-Specific Factor

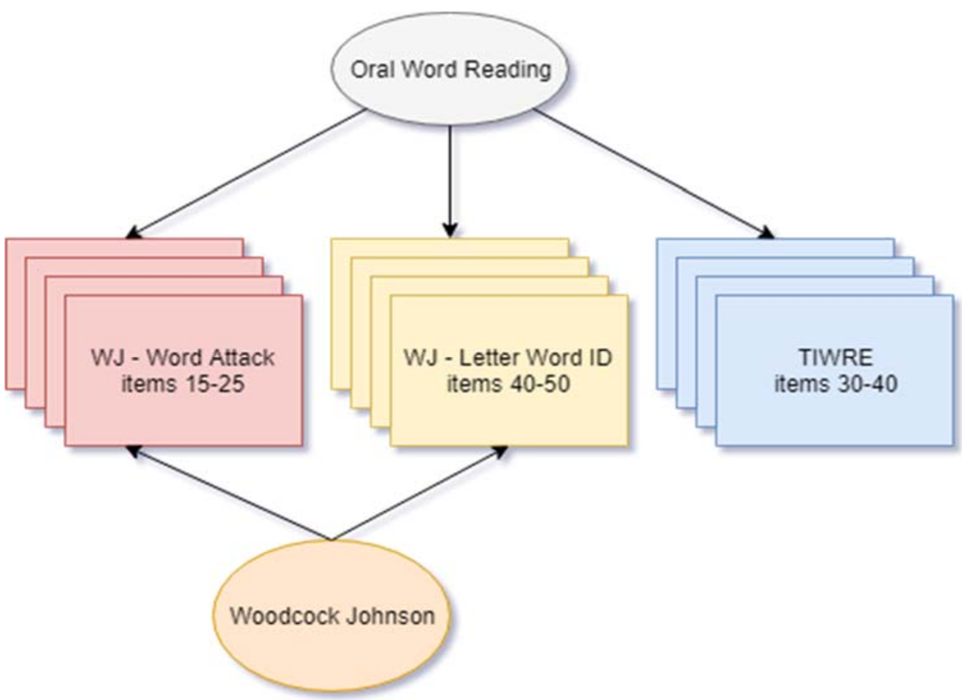

Figure 2.7. Bi-Factor Model With Test-Specific Factor

Research Question 2: To what extent might the measurement structure be equivalent in both native and non-native speakers of English?

To answer the question of measurement equivalence specified here, we started with single-factor models within tests and across tests, and the results indicated that when accounting for native speaker status through a specification of the configural model, with freely estimated loadings and thresholds by group, the overall model fit was improved for each within-test model (see Table 2.6). We then proceeded with the models from the full group analysis with the best fit, which were the bi-factor models with one overall across-test factor and an individual within-test factor for the shared trait (real words) and shared method (test battery). We found a continuous good fit for the sequence of baseline/configural, metric, and scalar models for both the trait and the method approach (see Tables 2.7 and 2.8). These sequences of models evaluated test structure (baseline/configural model), factor loadings (metric model), and then also the item thresholds (scalar model), to determine if the assessments had significantly different measurement properties across groups (Meredith, 1993; Vandenberg & Lance, 2000). In the trait-focused model, the WJ-LWI and the TIWRE subtests loaded onto the real-word factor, and in the method-focused model, the WJ-LWI and the WJ-WA subtests loaded onto the WJ shared factor. All items of all tests also loaded onto an overall word-reading factor. The model fit results under these two conditions were very similar (see Tables 2.7-2.14).

Table 2.6
Model Fit – Single-Test Models by Native Speaker Status

Models	CFI	TLI	RMSEA	SRMR	χ^2	df
WJ-Word Attack 1-Factor	.997	.996	.023	.044	108.805	88
WJ-Letter Word ID 1-Factor	.989	.986	.040	.064	149.723	88
TIWRE 1-Factor	.994	.992	.029	.064	107.433	88

Note: WJ=Woodcock Johnson, TIWRE=Test of Irregular Word Reading Efficiency The χ2 is calculated using the Satorra and Bentler mean-and-variance-adjusted (scaled) chi-square method (Satorra & Bentler, 2010).

Table 2.7
Model Fit - Bi-Factor - Method (Test Factor)

Model	CFI	TLI	RMSEA	SRMR	χ^2	Df	χ^2	df
Configural	.993	.993	.018	.078	1013	884		
Metric	.990	.989	.022	.094	1135	936	98*	52
Scalar	.986	.985	.025	.096	1242	966	155*	30

*Note: * indicates significance at p<.000 in a difference test comparison with the model above. Despite significant chi-squared comparisons across the three models, the differences in CFI and RMSEA are insignificant for measurement invariance purposes (Chen, 2007).*
The χ2 is calculated using the Satorra and Bentler mean-and-variance-adjusted (scaled) chi-square method (Satorra & Bentler, 2010).

Table 2.8
Model Fit - Bi-Factor - Trait (Real/Non-Word Factor)

Model	CFI	TLI	RMSEA	SRMR	χ^2	df	χ^2	df
Configural	.996	.995	.015	.075	970	886		
Metric	.990	.990	.021	.093	1122	937	142*	75
Scalar	.987	.986	.024	.094	1225	967	110*	44

*Note: * indicates significance at p<.000 in a difference test comparison with the model above. Despite significant chi-squared comparisons across the three models, the differences in CFI and RMSEA are insignificant for measurement invariance purposes (Chen, 2007).*
The χ2 is calculated using the Satorra and Bentler mean-and-variance-adjusted (scaled) chi-square method (Satorra & Bentler, 2010).

Table 2.9

Unstandardized Thresholds from the Trait Bi-Factor Scalar Model

Item	Threshold	SE
TIWRE30	-3.25	0.664
TIWRE31	-2.39	0.306
TIWRE32	-2.39	0.397
TIWRE33	-2.07	0.261
TIWRE34	-2.23	0.472
TIWRE35	-1.13	0.149
TIWRE36	-0.56	0.107
TIWRE37	-0.73	0.11
TIWRE38	-0.62	0.11
TIWRE39	0.21	0.11
TIWRE40	-2.51	0.294
WJLWI40	-2.24	0.224
WJLWI41	-1.22	0.090
WJLWI42	-1.27	0.090
WJLWI43	-1.34	0.134
WJLWI44	-0.49	0.096
WJLWI45	-0.31	0.106
WJLWI46	-2.03	0.198
WJLWI47	-0.47	0.088
WJLWI48	-0.74	0.086
WJLWI49	-0.54	0.093
WJLWI50	0.24	0.055
WJWA15	0.05	0.088
WJWA16	0.70	0.075
WJWA17	0.52	0.069
WJWA18	0.35	0.066
WJWA19	0.43	0.060
WJWA20	0.38	0.076
WJWA21	0.88	0.086
WJWA22	0.50	0.099
WJWA23	0.75	0.087
WJWA24	1.05	0.082
WJWA25	-3.25	0.664

Note: TIWRE=Test of Irregular Word Reading Efficiency, WJLWI=Woodcock Johnson Letter-Word Identification WJWA=Woodcock Johnson Word Attack

Table 2.10
Unstandardized Thresholds from the Test Bi-Factor Scalar Model

Item	Threshold	SE
TIWRE30	-2.71	0.365
TIWRE31	-2.50	0.352
TIWRE32	-1.85	0.197
TIWRE33	-1.85	0.178
TIWRE34	-1.37	0.141
TIWRE35	-1.08	0.133
TIWRE36	-0.55	0.100
TIWRE37	-0.55	0.080
TIWRE38	-0.68	0.112
TIWRE39	0.01	0.110
TIWRE40	-2.36	0.252
WJLWI40	-3.57	1.047
WJLWI41	-1.21	0.088
WJLWI42	-1.27	0.089
WJLWI43	-1.39	0.145
WJLWI44	-0.50	0.092
WJLWI45	-0.32	0.104
WJLWI46	-1.92	0.180
WJLWI47	-0.40	0.084
WJLWI48	-0.74	0.086
WJLWI49	-0.50	0.088
WJLWI50	0.24	0.060
WJWA15	-0.14	0.088
WJWA16	0.64	0.074
WJWA17	0.49	0.069
WJWA18	0.27	0.068
WJWA19	0.45	0.065
WJWA20	0.39	0.080
WJWA21	0.84	0.084
WJWA22	0.66	0.123
WJWA23	0.71	0.086
WJWA24	1.31	0.116
WJWA25	-2.71	0.365

Note: TIWRE=Test of Irregular Word Reading Efficiency, WJLWI=Woodcock Johnson Letter-Word Identification WJWA=Woodcock Johnson Word Attack

Table 2.11
Standardized Loadings from the Trait Bi-Factor Scalar Model

Item	Loading, General Factor, Native	Loading, General Factor, Nonnative	Loading, Trait-specific, Native	Loading, Trait-specific, Nonnative
TIWRE30	0.70	0.73	0.57	0.28
TIWRE31	0.79	0.75	0.38	0.17
TIWRE32	0.61	0.65	0.65	0.33
TIWRE33	0.65	0.64	0.55	0.26
TIWRE34	0.53	0.60	0.72	0.39
TIWRE35	0.71	0.66	0.44	0.20
TIWRE36	0.69	0.61	0.39	0.17
TIWRE37	0.48	0.43	0.53	0.23
TIWRE38	0.73	0.65	0.32	0.14
TIWRE39	0.81	0.71	0.14	0.06
TIWRE40	0.71	0.67	0.46	0.21
WJLWI40	0.58	0.53	0.48	0.21
WJLWI41	0.70	0.57	0.01	0.00
WJLWI42	0.67	0.55	0.05	0.02
WJLWI43	0.74	0.68	0.37	0.16
WJLWI44	0.82	0.72	0.13	0.05
WJLWI45	0.84	0.77	0.23	0.10
WJLWI46	0.77	0.70	0.35	0.15
WJLWI47	0.70	0.63	0.37	0.16
WJLWI48	0.68	0.58	0.32	0.13
WJLWI49	0.78	0.68	0.21	0.09
WJLWI50	0.59	0.47	0.57	0.28
WJWA15	0.84	0.74		
WJWA16	0.75	0.63		
WJWA17	0.73	0.61		
WJWA18	0.71	0.59		
WJWA19	0.65	0.52		
WJWA20	0.79	0.68		
WJWA21	0.80	0.69		
WJWA22	0.87	0.79		
WJWA23	0.82	0.72		
WJWA24	0.76	0.64		
WJWA25	0.70	0.73		

Note: TIWRE=Test of Irregular Word Reading Efficiency, WJLWI=Woodcock Johnson Letter-Word Identification WJWA=Woodcock Johnson

Table 2.12
Standardized Loadings from the Method Bi-Factor Scalar Model

Item	Loading, General Factor, Native	Loading, General Factor, Nonnative	Loading, Test-specific, Native	Loading, Test-specific, Nonnative
TIWRE30	0.86	0.76		
TIWRE31	0.88	0.79		
TIWRE32	0.84	0.72		
TIWRE33	0.83	0.71		
TIWRE34	0.78	0.65		
TIWRE35	0.82	0.71		
TIWRE36	0.78	0.65		
TIWRE37	0.63	0.48		
TIWRE38	0.81	0.69		
TIWRE39	0.84	0.73		
TIWRE40	0.82	0.71		
WJLWI40	0.83	0.68	0.03	0.05
WJLWI41	0.62	0.46	-0.32	-0.50
WJLWI42	0.63	0.48	0.29	0.41
WJLWI43	0.83	0.72	0.21	0.30
WJLWI44	0.78	0.64	0.02	0.03
WJLWI45	0.85	0.74	0.23	0.36
WJLWI46	0.82	0.70	0.17	0.27
WJLWI47	0.78	0.65	0.13	0.21
WJLWI48	0.74	0.61	0.08	0.12
WJLWI49	0.78	0.64	0.04	0.07
WJLWI50	0.53	0.38	0.20	0.30
WJWA15	0.81	0.67	0.26	0.35
WJWA16	0.69	0.53	0.18	0.28
WJWA17	0.66	0.50	0.25	0.37
WJWA18	0.67	0.51	0.27	0.38
WJWA19	0.57	0.41	0.20	0.29
WJWA20	0.70	0.53	0.30	0.41
WJWA21	0.72	0.57	0.35	0.50
WJWA22	0.77	0.58	0.29	0.43
WJWA23	0.74	0.58	0.45	0.64
WJWA24	0.63	0.45	0.31	0.46
WJWA25	0.86	0.76	0.47	0.63

Note: TIWRE=Test of Irregular Word Reading Efficiency, WJLWI=Woodcock Johnson Letter-Word Identification WJWA=Woodcock Johnson

The results of the bi-factor model informed by the Schmid and Leiman solution (1957) fit well for both the trait and method approach of structuring these models (trait CFI = .983 TLI = .981; RMSEA = .025; SRMR = .067; method CFI = .989 TLI = .987; RMSEA = .021; SRMR = .058; see Table 2.7 for further fit indices) when used for the full sample. In testing for equivalence across native and non-native speaking participants, the configural, metric, and scalar invariance models all fit quite well (Table 2.7 and Table 2.8). The configural bi-factor model with the loadings and intercepts loading freely indicated a good fit, and the metric model which restricted the loadings to be equal for both groups and scalar model which did the same for intercepts showed a significant but not substantial decrease in model fit (Chen, 2007). Towards the end of both the trait and test sequences, the SRMR was near the values recommended by Hu & Bentler (1995). However, as Marsh, Hau, and Wen (2004) point out in their evaluation of these commonly-used fit criteria, SRMR is particularly sensitive to complex data structure and sample size, so the evaluation in this study relied primarily on CFI, TLI, and RMSEA. Overall, while the model fit results for both sequences were very similar in all three stages of the measurement invariance testing, when comparing the final scalar models, the trait-focused structure was a better fit, indicating that the best model of the relations across these skills accounts for sight word recognition and phonic decoding.

The thresholds from both final scalar models indicated that the items were in most cases ordered by difficulty as expected for both native and non-native groups, aside from one item from the WJ Letter-Word ID and an overall more gradual increase for the phonological test (see Figure 2.8-2.9). Considering ordering of difficulty can be a way of identifying within-test and across-test differences across groups, and for these assessments, while some items were not in a completely linear sequence of increasing difficulty for these adults, we did not find specific patterns of item difficulty being out of order or unexpectedly clustered by difficulty. While the loadings for both scalar models were similar and in most cases statistically significant across models, the method model had more instances of the loadings being low or even negative and non-significant (see Table 2.11 and Table 2.12). Importantly, cases of low loadings in both models were consistent across native and non-native groups, further supporting the conclusion of measurement invariance across language groups. In the scalar model, the factor means were freed for the non-native group and constrained for the native group, in order to represent difference in the latent factor across groups (see Table 2.13 and Table 2.14). These results indicate that the non-native group outperformed the native group on the WJ subtests only, on the sample of items included in this study.

Table 2.13
Factor Means and Variances from the Trait Bi-Factor Scalar Model, Non-native

Factor	Mean	SE	Variance	SE
Overall word reading	.202*	.06	.509*	.07
Real word reading	-1.094*	.12	.117	.08

*Note: * indicates significance at p<.05. Mean and variance of the factors were set to 1 and 0, respectively in the native English speaking group.*

Table 2.14
Factor Means and Variances from the Method Bi-Factor Scalar Model, Non-native

Factor	Mean	SE	Variance	SE
Overall word reading	-.354*	.08	.477*	.08
Woodcock Johnson	1.245*	.17	1.682*	.46

*Note: * indicates significance at p<.05. Mean and variance of the factors were set to 1 and 0, respectively in the native English speaking group.*

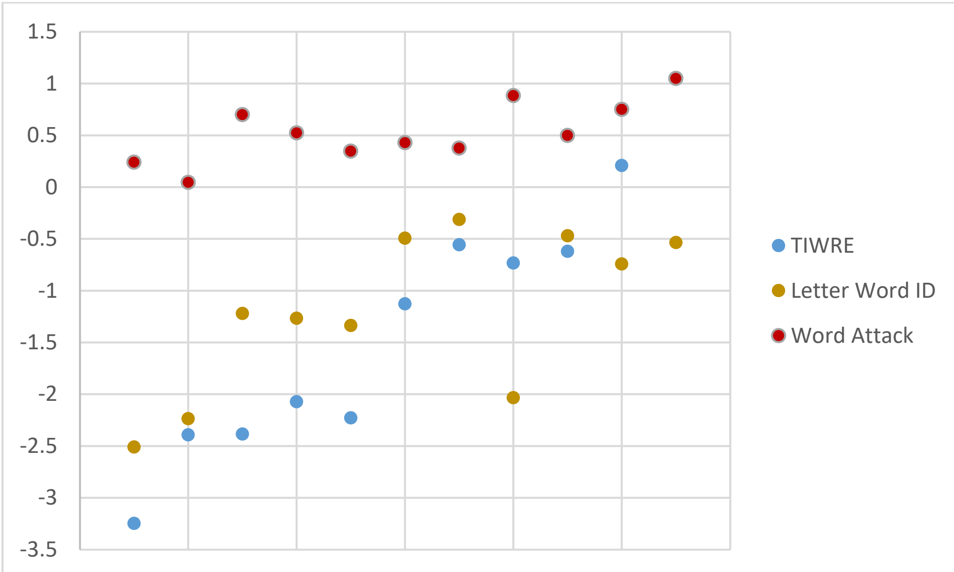

Figure 2.8. Unstandardized Item Thresholds – Trait Model

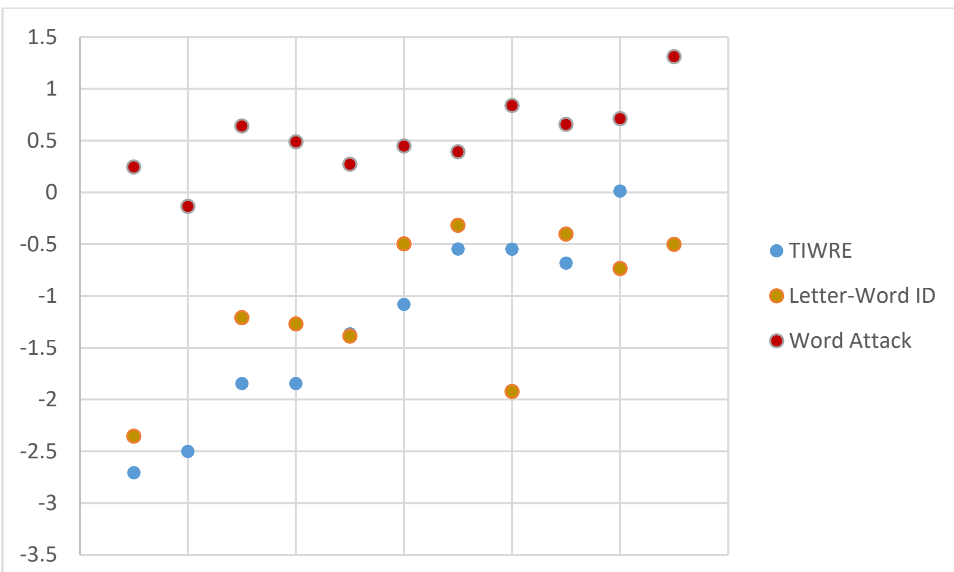

Figure 2.9. Unstandardized Item Thresholds – Method Model

To summarize, despite significant chi-squared comparisons across the three models, the differences in model fit which are seen in the CFI and RMSEA are insignificant for measurement invariance purposes (Chen, 2007), and we can determine from this that the trait-based bi-factor scalar model is the best fit from these three measurement invariance models, though the results for the trait and method (test) bi-factor models at the scalar level of measurement invariance testing had very similar model fit results. Model fit (see Table 2.7 and 2.8) indicate in most cases a good fit for both groups, although the trait factor seems to be a stronger influence for the native group, and the thresholds (see Figure 2.8-2.9) once again showed mostly consistent pattern of increase in difficulty. Examining the within-test structure of difficulty, the thresholds show a slight but consistent increase in item difficulty within the tests. Results of the full model sequence confirm measurement invariance across native/non-native speakers.

Discussion

This study examined the extent to which the items of these three tests measure real and non-word reading as compared to method-specific skills, and how that measurement structure might be equivalent in both native and non-native speakers of English, in this sample of struggling adult readers. Results of the full group analysis in this study indicate that across these three tests, there is one distinct overall word-reading factor measured across them, and that the test-specific method factors are separate and significant for both the native and non-native speaking groups. Full-group performance indicated that phonic decoding was an area of relative weakness, and when disaggregated by native speaker status, non-native speakers were found to have performed significantly better than native speakers on the measure of phonic decoding. These findings build upon previous research with this population which indicates that native and non-native speakers have differing strengths and weaknesses in the area of phonic decoding and orthographic word recognition, by further examining the item-level function of these assessments across groups (MacArthur et al, 2010; Davidson & Strucker, 2002). When assessing measurement invariance with native and non-native speakers, though the difference in fit between the configural, metric, and scalar model was statistically significant in each case, the differences were small enough that we may still consider the items invariant across groups

(Chen, 2007). The items have reasonably equivalent measurement properties for both native and non-native speakers of English.

Item-level outcomes

The item parameters suggest that the test items function reasonably well for measuring general word reading (loadings generally .6 or higher), and that their thresholds are generally increasing as would be expected to validate the implementation of basal and ceiling rules, with the non-word test (WJ-WA) collectively having the lowest loadings and most inconsistent order of difficulty. For researchers and practitioners, it seems reasonable to assume measurement equivalence across native and non-native speakers on these tests (i.e., there is no strong evidence of bias). Because there is no prior research comparting native and non-native speakers in this population on the item-level specifically within this group of learners, further research using this methodology with other assessments could continue answering the question of the item-level appropriateness of these measures for this population.

Group-level outcomes

Group differences vary across the two models. The trait model suggests that the non-native group did not perform as well as the native group on the real-word factor, though they outperformed the native group on overall word reading and non-word reading. The method model suggests that method may be more of an influential factor for the native group, and the oddities of this model indicate further analysis may be useful to determine if test method may not be equally as important across groups. This implies that non-native speakers may be more able to apply the rules of decoding and to recognize phonetically-regular non-words compared to native speakers of English, but that their overall lower familiarity of English results in lower orthographic word recognition and overall word reading performance. Based on previous research with this population which indicates that native and non-native speakers within this population may perform differently on reading assessments, and what we know of the relative strengths and weaknesses of each group, this is not surprising in terms of what each group finds more difficult, but this further elaborates on how that impacts the modeling of the relation of these skills (MacArthur et al, 2010; Davidson & Strucker, 2002).

Model outcomes

There is a strong general factor of word reading, but the trait model suggests that at least for non-natives, and perhaps for native speakers of English, non-word phonetic blends require additional processing on which people differ. This factor may not be equally important across groups. The method (test battery) used was also a significant factor, indicating that the method of assessment – or how similar in format tests are – may also be a factor affecting outcome. However, the trait model seems more theoretically straightforward. This builds on prior research which indicates that these skills may not be related in struggling adult readers as they are in children, by modeling this on the item-level and further examining the role native speaker status may play in this (Greenberg et al, 2011; Greenberg et al, 1997; Mellard et al, 2010; Nanda et al, 2010; Sabatini et al, 2010).

Limitations

An important limitation in this study is the lack of homogeneity within the population, as this sample of adult learners differed in age range, employment status, educational background, and more, all of which resulted in quite a varied sample. For the non-native group specifically, a wide variety of first languages were found as well (see Table 2.2), and generalizability within this group may be limited as a result. Across groups, the demographic representation was comparable, though a significant difference was observed in racial representation across native/non-native groups (see Table 2.2). This can make the interpretation of results difficult, as there may be complexities in the results that are not apparent from a full-group analysis or even with accounting for only some grouping variables like native speaker status. In addition to the variety of first languages represented in the non-native group, future research may want to closely consider background reading skills in one's native language, as drawing implications about the group of all non-native speakers may be complex in multiple ways. The grouping of native and non-native speakers here was of particular interest because of the potential implications on the model structure of word reading, but future research outside of the scope of this study on other subgroups in this population, based on demographic factors such as age range, gender, and education background, can be critical in order to ensure a valid application and interpretation of these assessments.

On the assessment side of this, the tests themselves present a limitation in that they are not necessarily designed for this population, which is an important part of why we are looking into this topic for this group. Additionally, because of the lack of prior research on these tests and with this group, these questions are exploratory, and there is therefore a limit of what we can expect or hypothesize. Because of the sample size and limited variance in responses to these items, the item-level models in this study are also limited to a subset of the words on this test. Similar item-level models with other measures, items, and subgroups may help us to better understand group and method-based differences.

Conclusions

Based on the previous research with this population which has indicated that compared to typically developing child readers, adult struggling readers' literacy skills are not necessarily as strongly correlated with each other, it is not surprising to see that the assessments studied here did not fit a joint single-factor model well (Greenberg et al, 2011; Greenberg et al, 1997; Mellard et al, 2010; Nanda et al, 2010; Sabatini et al, 2010). The item-level focus and MTMM approach to these questions of test function and population differences sets this research apart from other studies in this area. The method (test battery) and trait (real-word status) factors allowed us to see complexities in what these assessments actually measure with this sample. While the results indicated that these measures are invariant across groups, the means and variances taking into account native speaker status do reveal more about how the tests and items function with these groups, and there is still much to explore with examining how real/non-words may measure the skill of word reading differently in different language, skill, or age groups. The results of this study have built upon existing word reading research with struggling adult learners by taking a particular focus on the item-level information from these tests, while also presenting how native speaker status effects the model which best describes what these items and tests are actually measuring.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for Educational and Psychological Testing.* Washington, DC: AERA.

Alvermann, D. E., Unrau, N. J., & Ruddell, R. B. (Eds.). (2013). *Theoretical Models and Processes of Reading (6th Edition)*. Newark, DE, USA: International Reading Association.

Babayiğit, S. (2015). The relations between word reading, oral language, and reading comprehension in children who speak English as a first (L1) and second language (L2): a multigroup structural analysis. *Reading & Writing*, *28*(4), 527-544.

Binder, K., & Borecki, C. (2008). The use of phonological, orthographic, and contextual information during reading: a comparison of adults who are learning to read and skilled adult readers. *Reading and Writing*, *21*, 843-858.

Bosse, M., Chaves, N., Largy, P., & Valdois, S. (2015). Orthographic learning during reading: The role of whole-word visual processing. *Journal of Research in Reading, 38*(2), 141-158.

Bradley, L., & Bryant, P. (1983). Categorizing sounds and learning to read-a causal connection. *Nature*, *301*, 419–421.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, 56, 81-105.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504.

Cummings, K., Dewey, E., Latimer, R., & Good Iii, R. (2011). Pathways to word reading and decoding: The roles of automaticity and accuracy. *School Psychology Review*, *40*(2), 284.

Davidson R. & Strucker J. (2002) Patterns of word-recognition error among adult basic education native and non-native speakers of English. *Scientific Studies of Reading*. *6*(3):299–316.

Dunn, L.M., & Dunn. D.M. (2007). *Peabody Picture Vocabulary Test - Fourth Edition.* Boston, MA: Pearson.

Eason, S., Sabatini, J., Goldberg, L., Bruce, K., & Cutting, L. (2013): Examining the relationship between word reading efficiency and oral reading rate in predicting comprehension among different types of readers. *Scientific Studies of Reading*, *17*, 199-223.

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

Ehri, L. (2014). Orthographic mapping in the acquisition of sight word reading, spelling memory, and vocabulary learning. *Scientific Studies of Reading, 18*(1), 5-21. doi:10.1080/10888438.2013.819356

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., . . . Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*(1), 4-19. doi:10.1177/0146621606289485

Grainger, J., & Hannagan, T. (2014). What is special about orthographic processing? *Written Language & Literacy, 17*(2), 225-252. doi:10.1075/wll.17.2.03gra

Greenberg, D., Pae, H. K., Morris, R. D., Calhoon, M. B., & Nanda, A. O. (2009). Measuring adult literacy students' reading skills using the Gray Oral Reading Test. *Annals of Dyslexia*, *59*(2), 133–149. http://doi.org/10.1007/s11881-009-0027-8

Greenberg, D. Ehri, L., & Perin, (2002). Do adult literacy students make the same word-reading and spelling errors as children matched for word-reading age? *Scientific Studies of Reading*, *6*, 221–243.

Greenberg, D., Ehri, L. C., & Perin, D. (1997). Are word reading processes the same or different in adult literacy students and 3rd–5th graders matched for reading level? *Journal of Educational Psychology*, 89, 262–275. doi: 10.1037/0022-0663.89.2.262

Greenberg, D., Wise, J., Morris, R., Fredrick, L., Rodrigo, V., Nanda, A., & Pae, H. (2011). A randomized control study of instructional approaches for struggling adult readers. *Journal Of Research On Educational Effectiveness*, *4*(2), 101-117. doi:10.1080/19345747.2011.555288

Hannon, B. 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. doi:10.1002/RRQ.013

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

Joseph K., Torgesen, Wagner, R., & Rashotte, C. (1999). *Test of Word Reading Efficiency*. Austin, TX: PRO-ED.

Kilpatrick, D. A. (2015). *Essentials of assessing, preventing, and overcoming reading difficulties*. Hoboken, NJ: John Wiley & Sons.

Leinenger, M. (2014). Phonological coding during reading. *Psychological Bulletin, 140*(6), 1534-1555. doi:10.1037/a0037830

MacArthur, C. A., Konold, T. R., Glutting, J. J., Alamprese, J. A. (2010). Reading component skills of learners in adult basic education. *Journal of Learning Disabilities*, *43*(2), 108–121.

Mann, V. A. (1993). Phoneme awareness and future reading ability. *Journal of Learning Disabilities*, *26*, 259–269.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 320-341.

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology*, 4. doi:10.3389/fpsyg.2013.00169

McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). Technical Manual. *Woodcock Johnson IV.* Rolling Meadows, IL: Riverside.

McIntosh, K., Horner, R., Chard, D., Boland, J., & Good, R. (2006). The use of reading and behavior screening measures to predict non-response to schoolwide positive behavior support: A longitudinal analysis. *School Psychology Review*, *35*, 275–291.

Mellard, D. F., & Fall, E. (2012). Component model of reading comprehension for Adult Education participants. *Journal of Learning Disabilities*, *35*(1), 10-23.

Mellard, D., Fall, E., & Woods, K. (2010). A path analysis of reading comprehension for adults with low literacy. *Journal of Learning Disabilities*, *43*, 154–165. doi:10.1177/0022219409359345

Muthén, L. K., & Muthén, B. O. (1998-2018). *Mplus User's Guide*. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Nanda, A. O., Greenberg, D., & Morris, R. (2010). Modeling child-based theoretical reading constructs with struggling adult readers. *Journal of Learning Disabilities*, *43*, 139-153.

Nanda, A., Greenberg, D., & Morris, R. (2014). Reliability and validity of the CTOPP Elision and Blending Words subtests for struggling adult readers. *Reading and Writing: An Interdisciplinary Journal, 27*(9), 1603-1618.

National Research Council. (2012). *Improving adult literacy instruction: Options for practice and research* (A. M. Lesgold & M. Welch-Ross, Eds.). Washington, DC: National Academies Press

Nightingale, E., Greenberg, D., Branum-Martin, L. (2016). Selecting fluency assessments for adult learners. *Journal of Research and Practice for Adult Literacy, Secondary, and Basic Education.*

OECD Skills Outlook 2013: First results from the survey of adult skills, OECD Publishing. http://dx.doi.org/10.1787/9789264204256-en.

Pae, H., Greenberg, D. (2012). The relationship between receptive and expressive subskills of academic L2 proficiency in non-native speakers of English: A multigroup approach. *Reading Psychology, 35*(3), 221-259.

Paige, D., Rasinski, T., Magpuri-Lavell, T., & Smith, G. (2014). Interpreting the relationships among prosody, automaticity, accuracy, and silent reading comprehension in secondary students. *Journal of Literacy Research, 46*(2), 123-156

Rapp, D. N., Broek, P., van den, McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading, 11*(4), 289-312.

Reynolds, C. & Kampaus, R. (2007). *Test of Irregular Word Reading Efficiency*. Lutz, FL: Psychological Assessment Resources, Inc.

Sabatini, J., Sawaki, Y., Shore, J., & Scarborough, H. (2010). Relationships among reading skills of adults with low literacy. *Journal of Learning Disabilities*, *43*(2), 122-138

Satorra A & Bentler P. A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*. 2001; 66:507–514.

Shermis, M. D., & Long, S. K. (2009). Multitrait-multimethod analysis of FCAT reading and writing: Or is it writing and reading? *Journal of Psychoeducational Assessment*, *27*, 296–311. doi:10.1177/0734282908328639

Schmid, J. & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 83-90.

Thompkins, A. C., & Binder, K. S. (2003). A comparison of the factors affecting reading performance of functionally illiterate adults and children matched by reading level. *Reading Research Quarterly, 38*(2), 236-258.

Tighe, E. L., Schatschneider, C., Crepaldi, D., & Tomás, J. M. (2016). Modeling the relations among morphological awareness dimensions, vocabulary knowledge, and reading comprehension in adult basic education students. *Frontiers In Psychology*, *1-11*. doi:10.3389/fpsyg.2016.00086

Tighe, E. L., & Schatschneider, C. (2016). Examining the relationships of component reading skills to reading comprehension in struggling adult readers: A meta-analysis. *Journal of Learning Disabilities, 49*(4). 395-409.

Wagner R., Torgesen J., Rashotte C., Hecht S., Barker T., Burgess S., et al. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: a 5-year longitudinal study. *Developmental Psychology, 33* 468–478. 10.1037/0012-1649.33.3.468

Wagner, R., Torgesen, J., & Rashotte, C. (1994). Development of reading-related phonological processing abilities: New evidence of bidirectional causality from a latent variable longitudinal study. *Developmental Psychology*, *30*(1), 73–87.

Wiederholt, J., & Bryant, B. (2001). *Gray Oral Reading Tests.* Austin, Texas: Pro-ed.

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

Woodcock, R., McGrew, K., & Mather, N. (2001b). *Woodcock-Johnson III Tests of Cognitive Abilities*. Itasca, IL: Riverside Publishing

Appendix

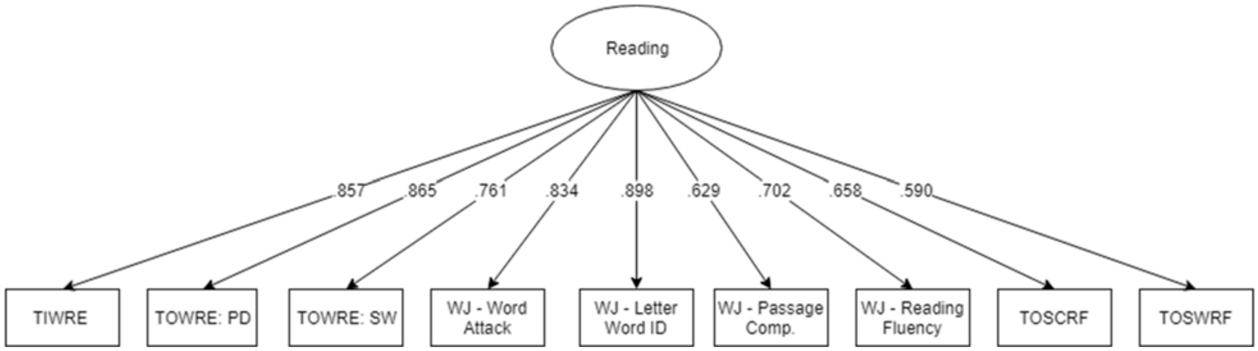

Figure 1.1: Single Factor

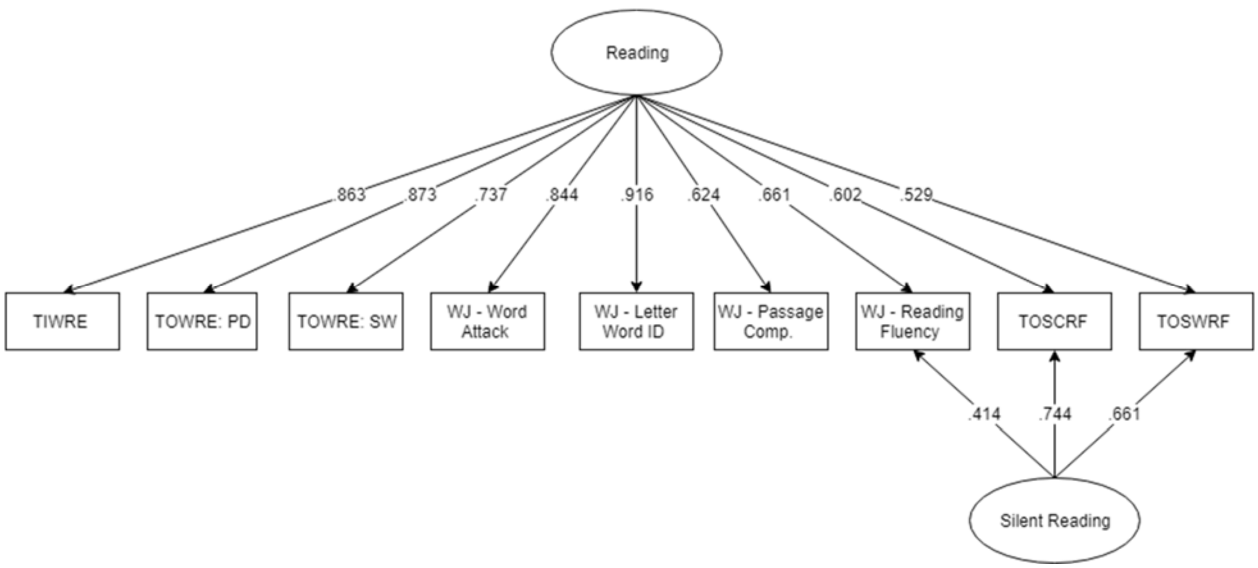

Figure 1.2: Silent Factor

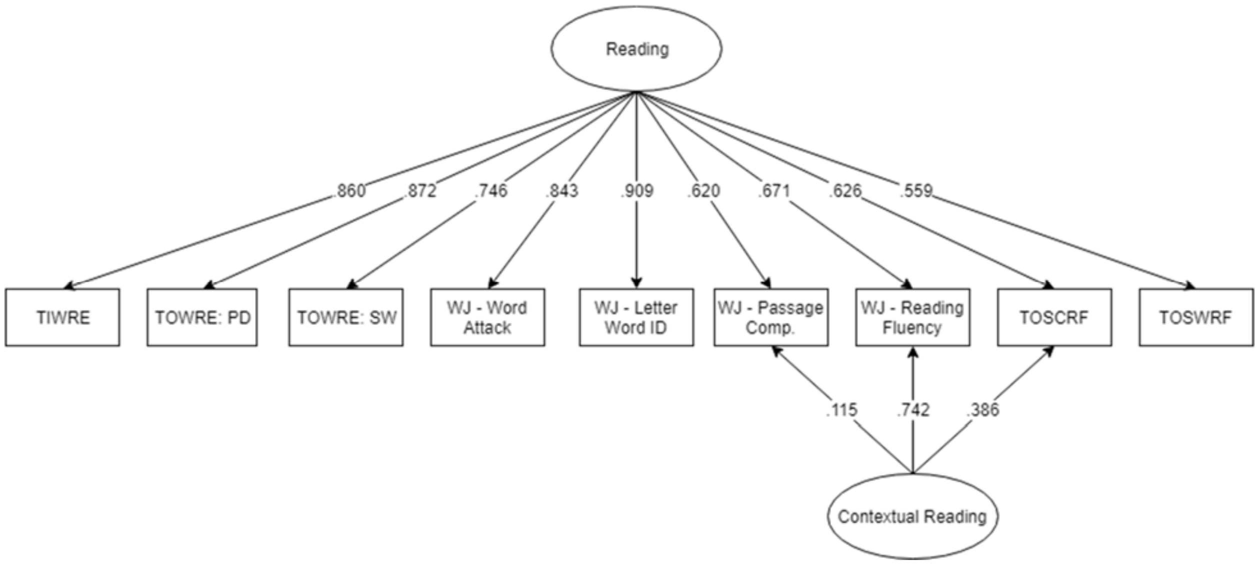

Figure 1.3: Contextual factor

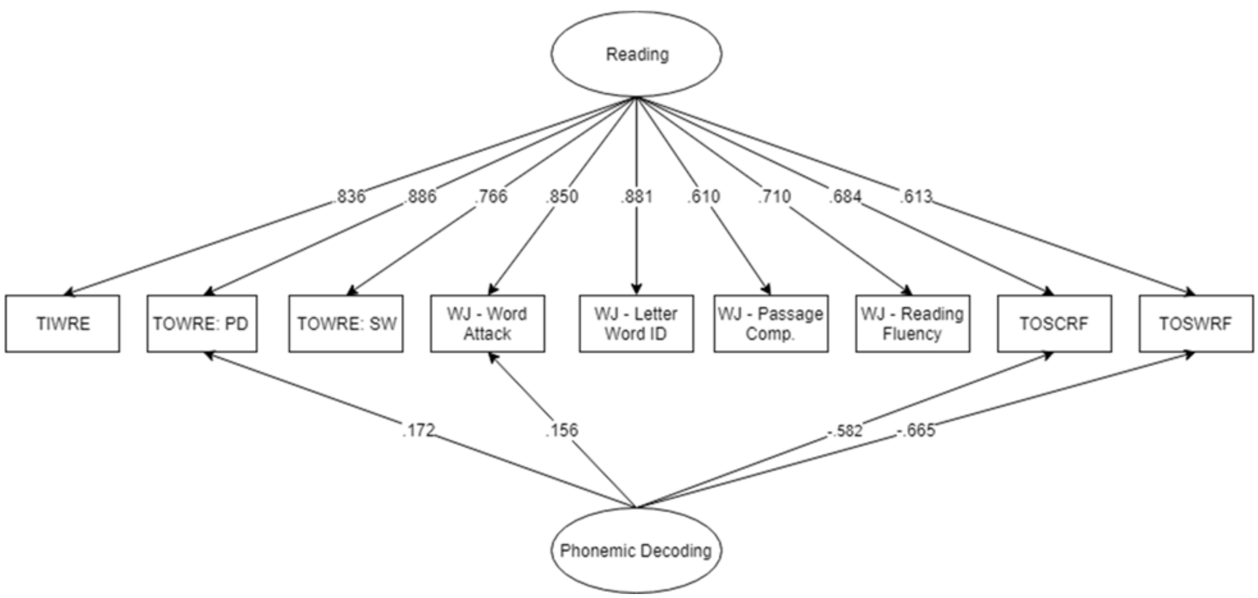

Figure 1.4: Phonemic Decoding Factor

Figure 1.5: Speeded Factor

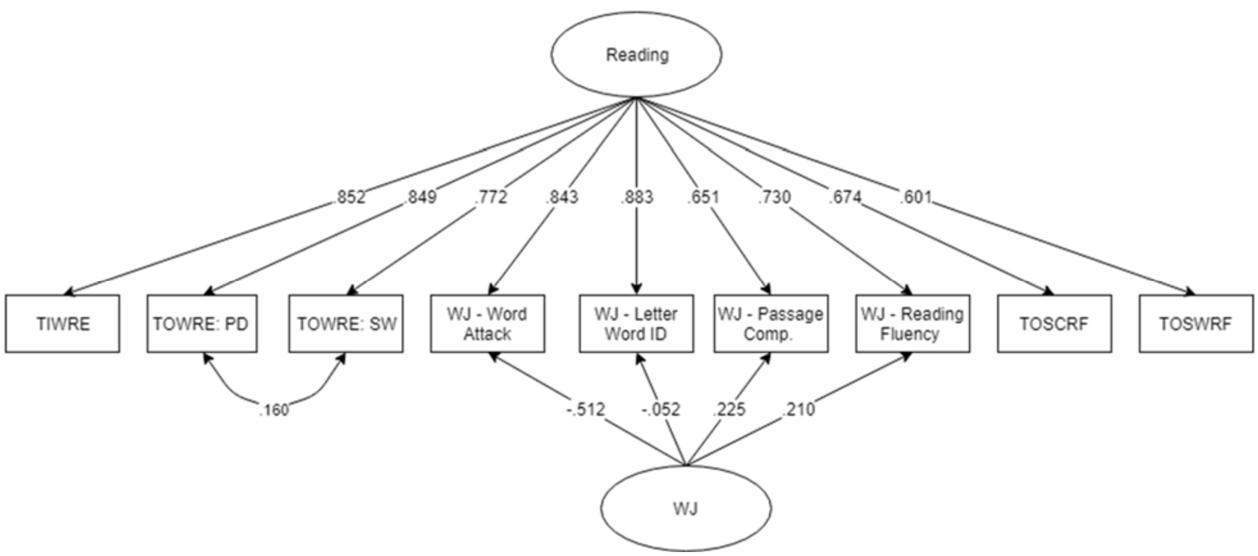

Figure 1.6a: Battery Factor

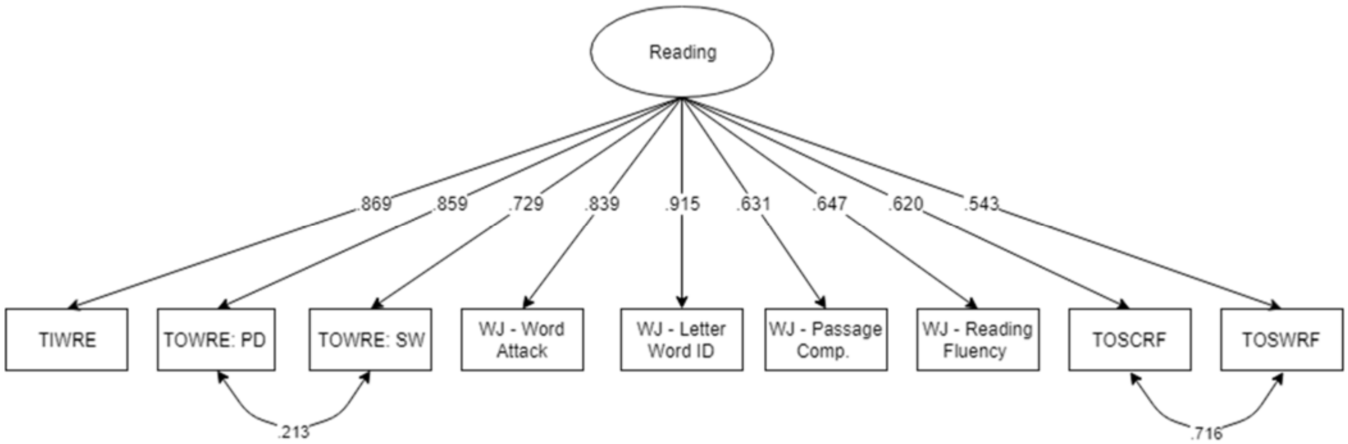

Figure 1.6b: Battery Factor

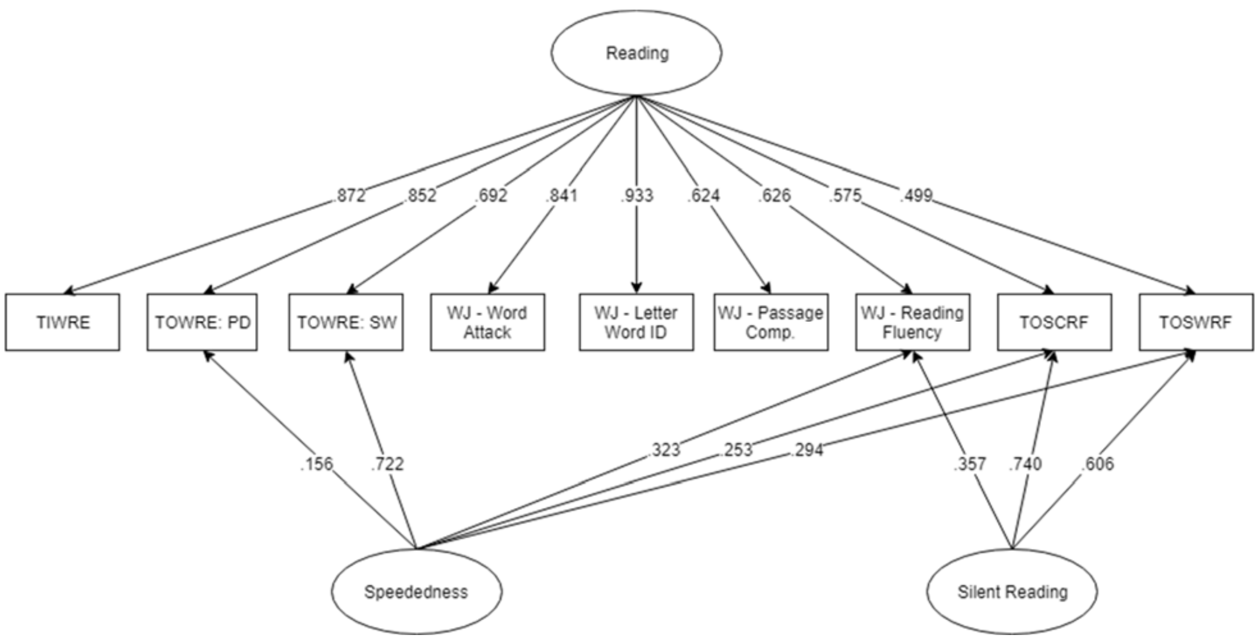

Figure 1.7: Method/Trait Factors