University of Massachusetts Amherst

ScholarWorks@UMass Amherst

Doctoral Dissertations

Dissertations and Theses

July 2017

Maximizing Test Efficiency: The Effects of Test Format on Nonsense Word Reading

Amanda Kern

MAXIMIZING TEST EFFICIENCY:

THE EFFECTS OF TEST FORMAT ON NONSENSE WORD READING

A Dissertation Presented

by

AMANDA M. KERN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2017

College of Education

MAXIMIZING TEST EFFICIENCY:

THE EFFECTS OF TEST FORMAT ON NONSENSE WORD READING

A Dissertation Presented

By

AMANDA M. KERN

Approved as to style and content by:

Michelle K. Hosp, Chair

John L. Hosp, Member

Amanda M. Marcotte, Member

Jill Hoover, Member

Joseph B. Berger, Senior Associate Dean
College of Education

DEDICATION

To Drew

Who always knew I could do this

ACKNOWLEDGMENTS

I can honestly say that this would not have been possible without an extraordinary group of people supporting me. This took a village, and my village is pretty awesome!

First, I would like to thank the members of my committee for their time and assistance. Michelle Hosp, my advisor and chair, thank you for your mentorship, for setting the bar high, for believing in me, and helping me through (literally) everything. Your friendship is something I will cherish forever. Next, John Hosp, the person responsible for getting me into this in the first place! Thank you for your guidance, and for bringing a much needed sense of humor to this process. Amanda Marcotte, thank you for taking in an orphan special education student and giving me a community that certainly contributed to my success. Jill Hoover, thank you for providing valued feedback and pushing me to consider the unique needs of every student.

A special thank you to my parents, Bob and Kathy Metcalf. Their support and love assured me that taking this on was the right choice. To the rest of the Metcalf and Kern family, especially Jason, Katie, Janis, Jeff, Alex, and Max: thank you for all the encouragement. To my friends and cohort members in Iowa and Massachusetts, thank you for all the support and help. Ashley: thanks for being both the greatest friend and my virtual therapist.

I am very lucky to have a partner that believed in me, and did not hesitate to support me during this journey, even when it involved moving across the country. Drew-I hope that I have made you proud. Thank you for being my best friend, and pushing me to do this in the beginning. I love you the most! And to Bella, who witnessed this dissertation from start to finish!

ABSTRACT

MAXIMIZING TEST EFFICIENCY: THE EFFECTS OF TEST FORMAT ON NONSENSE WORD READING

MAY 2017

AMANDA M. KERN, B.A., THE UNIVERSITY OF IOWA

M.A., THE UNIVERSITY OF IOWA

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Michelle K. Hosp

A repeated measures study was conducted to determine the effects of test format on accuracy and fluency performance on a computer-based, nonsense word, decoding task. Decoding is a phonics skill that is highly predictive of overall reading performance (Fletcher, Lyon, Fuchs, & Barnes, 2007). Therefore, identifying students who are struggling with decoding and providing instruction to remedy skill deficits is of high importance to teachers. A possible way for teachers to determine the instructional needs of their students is through testing (Hosp & Ardoin, 2008). However, time dedicated to test completion in classrooms limits the time available for instruction. Therefore, it is prudent that testing practices are efficient, but still yield reliable and valid data that can be used to inform instructional decision-making. This study examined how test format may be a variable that might improve the efficiency of decoding tests.

Fifty-three second grade students from a single elementary school in the northeast participated in this study. Participants completed a battery of decoding and reading tests. These included: A computer-based modified 100-word Nonsense Word Fluency (NWF) task that was formatted five ways, the DIBELS Next NWF benchmark, the Decoding

Inventory for Instructional Planning - Screener (DIIP-S), DIBELS Next Oral Reading Fluency (ORF) benchmark, and the Group Reading and Diagnostic Evaluation (GRADE).

Results from a series of repeated measures ANOVAs showed there are performance differences across test formats for both accuracy and fluency performance metrics. In addition, results show there are no performance differences across formats between student indicated preferred and non-preferred formats. Last, correlational analyses show there is evidence of criterion-related validity for each test format, but the strength of the evidence is dependent on test format, score metric, and criterion of interest.

The effect of test format on student performance indicates test format is a potential variable in exploring ways to improve the efficiency of decoding testing practices. Results align with previous research on the effects of the number of words presented at a time affecting reading speed, as fluency scores were significantly higher on the formats with more words, but diverge from previous research on the effects of student motivation, as results in this study did not find any effect on student preference for format. Implications for test development and directions for future research are discussed.

TABLE OF CONTENTS

Page

APPENDICES

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Reading as a Critical Skill

Learning to read is arguably the most critical goal of schooling. Reading failure, even in the early years of elementary instruction can be indicative of a variety of future problems. One in six third graders who do not meet reading proficiency benchmarks fail to graduate from high school on time. To be a proficient reader, one must be able to read a variety of texts with adequate fluency (i.e., speed and accuracy) as this allows them to focus on understanding what they are reading (Whitaker, Gambrell, & Mandel Morrow, 2004). Twenty-three percent of readers who fall below the reading benchmark in elementary school never finish high school. By comparison, only 4% of proficient readers do not finish high school (Hernandez, 2011). In addition to the negative academic outcomes, failure to achieve reading proficiency is associated with various negative long-term outcomes. They include: decreases in quality of life, decreased employability, economic security problems, and inability to live independently (Roman, 2004). An individual who does not acquire basic reading skills has difficulty interacting with print and text in meaningful ways, thus negatively affecting their quality of life. In addition, an individual who cannot read faces difficulty in acquiring and retaining meaningful employment, compromising their economic security (McCardle & Chhabra, 2004).

With such high-stakes for reading skill acquisition, there has been an increase in urgency in promoting effective instructional practices as a means of improving reading skills (Torgesen, 2002). National educational policies (e.g., No Child Left Behind Act (NCLB) in 2001, and the Every Student Succeeds Act (ESSA) in 2015) directly address requirements for reading proficiency. Special congressional commissions (e.g., the National Reading Panel (NRP)) have been charged with identifying the components of effective reading instruction and disseminating information to the schools. Federally funded reading programs, such as Reading First (included in NCLB), and most recently the Literacy Education for All, Results for the Nation (LEARN) program (included in ESSA) are designed to promote best-practices in reading education across the country. The LEARN program is intended to promote literacy skills for students at all levels of development beginning at birth and continuing through high school. In addition, explosions in reading research, with over 100,000 studies pertaining to reading published since the 1960s, show that the educational climate is primed for recommendations that can help improve student success in reading (National Reading Panel, 2000). The National Assessment for Educational Progress (NAEP) publishes the Nation's Report Card every two years. The Nation's Report Card includes information on the performance of students in grades 4, 8, and 12 across a variety of subject areas, including reading. It includes data used to determine the overall reading success of students in the United States, and is the gauge for measuring the success of national trends in reading.

Statistics on Reading Achievement

Despite the problems associated with failure to attain reading proficiency, and the implementation of various educational policies, reading achievement continues to be an area in need of improvement in schools. The NAEP describes a proficient fourth grade reader as one that can "integrate and interpret texts and apply their understanding of the text to draw conclusions" (NCES, 2015, NAEP Reading Achievement Levels, para. 5). In contrast, a fourth grade reader who scores in the basic level can "locate relevant information, make simple inferences, and use their understanding of the text to identify details that support a given interpretation" (NCES, 2015, Reading Achievement Levels, para. 4). In 2015, only 36% of fourth-grade students scored at or above the Proficient level in reading on the NAEP. Thirty-three percent of fourth-grade students scored in the Basic level, and 31% scored below Basic. These scores were not significantly different than the fourth-grade NAEP reading results from 2013 (NCES, 2015). In 2015, 34% of eighth-grade students scored at or above the Proficient level in reading on the NAEP. Forty-two percent of eighth-grade students scored in the Basic level, and 24% scored below basic. With a significant portion of students continuing to fail to meet reading proficiency standards it has become crucial that instructional practices alter to remediate the reading achievement problem. Proactive instructional practices that catch struggling students early, and intensive and effective interventions that remediate reading problems are likely to improve the outcomes for struggling readers. The question becomes, how do teachers identify those students that are in need of additional support? The simple answer is: they test them.

The Role of Decoding in Reading

When prioritizing which skills to test it is beneficial to select skills that are predictive of reading success. This improves the efficiency of testing and increases the time spent on instruction, rather than spent on testing. One skill that is highly predictive of reading success is decoding (Fletcher, Lyon, Fuchs, & Barnes, 2007). Decoding refers to a reader's ability to apply letter-sound correspondences to pronounce words (i.e., decoding is the skill that assists readers with word identification), and is included in the phonics skills category. Readers who experience difficulty with decoding are likely to have depressed overall reading skills, as accurate word identification is necessary for reading comprehension (i.e., the goal of reading) (Perfetti, 1986; Stanovich, 1991). In addition, decoding skill deficits are the most common problems for students who struggle with reading (Moats & Tolman, 2009). Therefore, when assessing students who are struggling, it is likely that the source of their problem resides at the word level. Testing for decoding skills will provide teachers with information that can inform their instructional planning.

Decoding is not the equivalent of reading. Instead, decoding is just one of several skills that is necessary for reading to occur. The report from the NRP identifies five skills that are necessary for reading: phonemic awareness (PA), phonics, vocabulary, fluency, and comprehension. PA includes a variety of auditory skills (e.g., phoneme blending and segmenting) that support students understanding of the language sounds that are used to build words. PA skills are critical to support young readers as they transition to identifying words in print. PA is necessary for the development of phonics skills. Once PA skills are developed, phonics skills are able to emerge. Phonics links what readers hear (using their PA skills) with what they see in print and includes several skills. Decoding is a specific phonics skill that refers to a reader’s ability to apply their letter-sound correspondences to pronounce unknown words. Phonics is necessary for proficient reading and is often the stumbling block for young readers. Readers who do not have proficient phonics skills often struggle with the other three critical skill areas necessary for reading (i.e., fluency, vocabulary, and comprehension). Phonics is necessary for developing adequate fluency, building vocabulary, and increasing reading comprehension abilities, as it is the skill that allows the reader to first identify the words in print. In this critical role it is pertinent that the tests teachers use to determine phonics skill proficiency provide accurate and useful results.

The Role of Testing in Improving Reading

While national educational policies emphasize the importance of testing as a means of quantifying broad educational achievement, classroom assessments can serve a range of educational purposes, and can positively impact student achievement (Hosp & Ardoin, 2008). The conventional approach to classroom assessment was mastery measurement (Jenkins & Fuchs, 2012). In mastery measurement, tests are given following instruction, and teachers use results to determine if a student has mastered a specific skill. The work of Stan Deno (1985) highlighted the technical problems and limitations of mastery measurement, including the impracticality of measuring every skill, the lack of generalization of skills measured, and technical inadequacies of mastery measures. In addition, the work of Deno and Phyllis Mirkin (1977) on Data-Based Program Modification as an alternate paradigm for classroom assessment highlighted the need to link instruction and assessment.

The shift to using assessment data to inform instruction, instead of using instruction to inform test development was critical to teacher’s understanding that tests could serve a variety of purposes in their classrooms. The purpose of the test dictates what test should be selected for use. As a screening tool, tests can be used to identify students at-risk of academic failure, before failure has occurred allowing for proactive intervention to take place. In a progress-monitoring role, tests can track student growth over time allowing teachers to implement instructional strategies that can increase the rate of student improvement. For diagnostic purposes, tests can identify specific skill deficits that can inform instructional practice by allowing teachers to target specific skill deficits without having to infer or assume what a student is struggling with. For outcome purposes, tests measure what a student has learned in a particular academic area (Hosp, Hosp & Howell, 2016). For decoding, the most common screening and progress monitoring tests are Nonsense Word Fluency (NWF) curriculum-based measures. An example of a diagnostic test for decoding is the Test of Word Reading Efficiency (TOWRE 2). In addition, decoding is included on several reading outcome measures such as the Woodcock Reading Mastery Test-Revised (WRMT-R). The tests that are available for teachers to use to measure decoding are dependent on the policies and resources available to them.

Limitations in Tests Available

The decisions on which tests teachers use are often influenced by the available resources, and policies of the school. State or district testing requirements dictate which broad or high-stakes accountability tests (e.g., Measures of Academic Progress (MAP) tests, Partnership for Assessment of Readiness for College and Careers (PARCC)) students will participate in each year. In the classroom, teachers are restricted to using the tests that the curriculum requires, that are required per school policy, or the physical testing materials that they have access to. Reliance on traditional testing products and procedures inhibits change in assessment practices, and limits the impact testing has on student outcomes (Box, Skoog, & Dabbs, 2015). However, improved assessment practices (e.g., adopting new or improved testing products, implementing universal screening procedures, or progress monitoring students in interventions) are positively associated with improved student outcomes (Black and Wiliam, 2009). Instructional decisions that are based on data are more likely to be accurate than decisions made without testing (Shepard, Hammerness, Darling-Hammond & Rust, 2005). In addition, teachers acknowledge and value the information that classroom assessments provide that can be applied to instruction and used to facilitate student learning (Black, 2014; Tobin, 2008). With positive associations with student outcomes, and positive teacher reception, testing improvements are likely to be well received by the field.

Test Format and Student Performance

The format of the test is one variable to consider when selecting a test. To date there have been few studies that have directly examined the relation between test format and student performance. Tests should be appropriately formatted for efficient administration and participation to decrease the instructional time requirements for completing the test. Classroom time spent on testing is time not spent on instruction. The presentation of test items, on reading tests, to students has remained largely unchanged over time. However, attention to how word layout and number of items presented at a given time affect student performance should be taken into consideration as these factors may impact their performance on the test. In addition, the increase in prevalence of technology in classrooms presents a possible alternative platform (i.e., using a computer for test administration) for testing that may improve testing efficiency for teachers.

An additional factor that may influence student performance on a test may be the student's preference for layout (e.g., words in columns versus words in rows) and number of items (e.g., 5 items versus two items). Students are likely to have increased motivation to participate in a task that accommodates their preference in format, or allows them to select the design of the task (Randi & Corno, 2000). Student motivation is an important factor to consider in reading (Morgan & Fuchs, 2007). Therefore, it should be considered in test development and design. As student motivation increases, performance on the task is likely to change when compared to performance of the same task presented in a non-preferred manner.

It is critical to maximize student's performance on a test when results are going to be used to guide and inform instruction. Test data are used to determine what skills should be included in instruction, and the effects of instruction are maximized when the instruction directly matches student instructional needs. Test results that do not accurately represent student need are unlikely to positively impact instructional planning for a student. In addition, it is prudent to examine ways that the test might be constructed to promote optimum student performance while minimizing the time required for administration.

Purpose of the Current Study

The purpose of this study is to determine the effects of test format on student decoding performance. Participants in this study read nonsense words presented in a variety of formats. Comparisons across formats will determine if there is an optimal format for accuracy and completion time. Comparisons of performance on the different test formats to other measures of reading will determine the how the effects of test format relate to broader measures of reading performance.

CHAPTER 2

LITERATURE REVIEW

Chapter Overview

Given the increased prevalence of testing in schools much of the research has been focused on how assessment practices impact student-learning outcomes. This chapter will address the research being done on the role of assessment as it pertains to decoding skills. Specifically, this chapter will review the research related to three areas. First, the theory and role of decoding in the broader context of reading performance will be reviewed. Second, the research on the role and importance, features, and history of decoding assessment practices will be shared. Last, the research on test format, and the role of student motivation in reading performance will be reviewed. The research reviewed in this chapter will provide a rationale for the current study examining test format as a means of improving testing efficiency in decoding assessments.

Definition of Terms

The following terms will be used throughout this study:

Assessment: a process of collecting information

Decoding: using letter-sound correspondences and word knowledge to convert printed text to spoken language

Grapheme: smallest unit of written language that represents sounds in words

Orthography: a system of written language that includes letter formations and spelling

Phoneme: smallest unit of sound in language

Phonics: the systematic process of teaching a person to read by using the sound-symbol connections in words

Phonology: the sound system of a language

Test: a specific measurement tool that quantifies student performance so that comparisons can be made between students

The Role of Decoding in Reading

Phonics is a broad component skill of reading. Phonics refers to an instructional approach in reading, and includes instruction on a variety of component skills. Decoding is one phonics skill that requires the reader to apply phoneme-grapheme knowledge to pronounce words (e.g., applying and blending the individual phonemes: /b//a//t/ to read /bat/). It requires the coordination and application of phonological knowledge to printed text. Within the context of general reading skill, decoding serves a critical role, which is providing readers with the ability to pronounce or identify words in text. Its function is well documented in several theories of reading, and its development has been studied through reading research.

Decoding in Reading Theory

Decoding is included in several reading theories. In the Simple View of Reading (SVR), decoding is identified as one of two critical skills that contribute to reading comprehension. The SVR states that reading comprehension is the result of the interaction between decoding and language comprehension (Gough & Tumner, 1986). Decoding is the ability for readers to recognize words, while language comprehension is the ability for readers to connect meaning to the words that they recognize. Together, when applied to written text, decoding and language comprehension interact to result in reading comprehension (i.e., making meaning from text). However, decoding skills alone do not supply the reader with enough skills to result in reading (i.e., the language comprehension piece must also be intact). Readers that have difficulty identifying words, but have intact language skills are often labeled dyslexic. Readers who can decode well, but struggle with associating meaning with the words read are often labeled hyperlexic. The SVR theory identifies that decoding is necessary, but not sufficient for reading.

Decoding is also identified as a critical component in the dual-route theory of reading. According to the dual-route theory readers identify words in two distinct ways (Forster & Chambers, 1973; Coltheart, 2005; Pritchard, Coltheart, Palethorpe & Castles, 2012). First, words that are familiar (i.e., immediately recognized) are recognized by sight and are immediate retrieved as whole units, from the reader's lexicon. This route of reading does not require the access or application of phonological skills to read the word. Rather, word identification is completed as a memory task. In contrast, words that are unfamiliar (i.e., not recognized by sight) are identified through a phonological process where the reader uses decoding skills to decipher the text to identify the unknown word. Dual-route reading theory maintains that decoding serves a critical function in reading.

In *Beginning to Read* (1990), Marilyn Adams described a connectionist approach to reading. According to the connectionist theory, reading is the result of the coordination and synthesis of a variety of processes. The processes Adams identified included: orthography, phonology, meaning, and context. Decoding is the application of orthographic and phonological processes to text. The stronger the connection between these processes the greater the likelihood that a reader will experience success (Adams, 1990). Decoding skills alone are not sufficient to result in success in reading, instead, just as the NRP reported, they are one of several necessary components that combine to result in proficient reading.

Last, decoding is a noted skill in automaticity theories of reading. The theory of automaticity holds that readers who are able to efficiently identify words in print are able to dedicate more of their cognitive resources to higher-level reading skills. In contrast, readers who struggle to efficiently identify words (i.e., have poor decoding skills) have limited cognitive resources available for comprehension or vocabulary building (Perfetti, 1984). Struggling readers extend so much effort at the word level, that they have difficulty connecting that word to any context. By comparison, skilled readers read with automatic word recognition abilities, allowing them to make connections beyond the word level (Gentry, 2006). Decoding skills promote efficient and accurate word identification. The theory of automaticity requires decoding skills be intact in order to support proficient reading.

Decoding is included in several reading theories. Each theory notes that decoding plays a critical role in general reading processes. However, none of the theories considers decoding to equal reading. Instead, decoding is noted as a necessary component within the reading process that when combined with other components (e.g., language comprehension, vocabulary, etc.) results in reading. Additional evidence of the role of decoding, and its function with other reading skills, are included in the NRP report. In isolation, decoding is not substantive enough to equate to reading. Instead it is the successful integration of decoding skills, that are supported by PA skills, and combined with other critical component skills (e.g., fluency, vocabulary, and comprehension) that results in successful reading. Decoding is necessary, but not sufficient for reading. In such a critical role, breakdowns in decoding skills can have significant effects on a reader’s ability to interact with text in meaningful ways. Understanding the developmental process of decoding skills will assist in timely identification of skill deficits that may impact overall reading development.

Development of Decoding

Decoding typically develops over time with readers first mastering simple (i.e., one-to-one) phoneme-grapheme correspondences and advancing to mastery of more complex spelling patterns in order to read words (e.g., affixes, digraphs, trigraphs, etc.) (Chall, 1996; Ehri, 1997). In the primary stage of word reading (i.e., pre-alphabetic or pre-reading phase) readers do not apply phonological awareness skills to read words. Instead, during this stage of development readers rely on visual cues to identify words. That is, readers associate some visual feature of the printed text with the word meaning. For example, readers might identify the word “look” by associating the “oo” to represent eyes. During this stage of development instruction is focused on developing phonological and phonemic awareness skills that bring student attention to the sound structure of language, and how sounds are manipulated to form words. These skills are necessary to transition to the next stage of word reading development.

It is during the second stage of word reading development (i.e., partial alphabetic or initial reading stage) that the first decoding skills emerge (Chall, 1996; Ehri, 1998). Intact phonemic awareness skills are critical for the transition from the first to second stage. During this stage readers form the connection between letters and sounds, and begin to apply those sounds to identify words in print (i.e., onset of decoding). During this stage readers may not apply letter-sound correspondences to all letters in a word. For example, it is common that readers in this stage to apply sounds to the first letter only in a word (e.g., reading “bat” as ball) followed by the last letter (e.g., reading “bat” as bit) and omitting medial sounds. Phonics instruction during this phase of decoding development is focused on accurate letter-sound associations, application of those sounds in single syllable words, and incorporation of mastered phonemic awareness skills (blending and segmenting) to connecting sounds in printed words to pronounce whole words. Once primary skills are mastered in the second stage, readers transition to the third stage of word reading development.

The third stage (i.e., full alphabetic stage) of word reading development involves readers applying known letter-sound correspondences in words, and learning additional and more complex letter patterns to assist in identifying words in print (Ehri, 1997). In this stage readers attend to all letters in a word to identify it, and accurately apply letter-sound correspondences to an entire word. For example, readers may read sound-by-sound, pronouncing /p/ /i/ /g/ to read /pig/. During this stage of development phonics instructional time is spent on expanding decoding knowledge to include more complex letter patterns (e.g., digraphs, blends), and providing students with ample opportunities to use their decoding skills when reading. Transition to the next stage of word reading development occurs as students increase their reading fluency and begin to segment words into multi-letter blocks to read more efficiently.

The fourth stage (i.e., consolidated alphabetic or fluency stage) of word reading development involves readers increasing their speed, and accuracy of identifying words (Chall, 1996; Ehri 1997). This requires students to maximize efficiency in applying decoding skills. To maximize efficiency readers use advanced decoding skills, such as rapid application of syllable rules, identifying root words, analogizing unknown words to known words (i.e., associating letter patterns in unknown words to words that they know), and breaking words into larger segments that can more quickly be identified and blended together to pronounce the unknown word. For example, readers may pronounce three segments: /per/ /fect/ /ly/ to read /perfectly/ by identifying the prefix /per/, the middle syllable as a closed syllable to assist with pronouncing the vowel, and the suffix /ly/. Phonics instruction during this stage is focused on those advanced decoding skills, and includes exposure to a variety of texts that allow the reader to practice those skills. It is critical that decoding skills are intact, as breakdowns or skill deficits can negatively impact success in reading.

The Impact of Decoding

For most struggling readers, the source of their reading difficulty resides at the word level (Knutson, Simmons, Good & McDonagh, 2004; Shaywitz, 2003; Stanovich, 1991; Torgesen, 2000). Up to 80% of students with a specific learning disability in reading struggle with reading at the word level (Moats & Tolman, 2009). The ability to accurately and efficiently read words impacts broader reading development and word reading abilities are highly correlated with overall reading ability (Fuchs, Fuchs, Hosp, & Jenkins, 2001). While accurate word reading does not guarantee reading comprehension will occur, reading comprehension is impossible without intact word reading skills. Readers who devote complete attention at the word, or even letter level have few cognitive resources left for connecting meaning to what they have read (Stanovich, 1981). Problems with reading comprehension, or reading fluency can most often be traced back to breakdowns in a reader’s word reading skills (Carver, 1998; Murray, Munger, & Clonan, 2012; Perfetti, 1986).

The skill breakdown for many students that experience difficulty with reading is at the word level where they have difficulty pronouncing individual words in text. Readers who are stuck at the word level will have difficulty with reading fluency and will have difficulty comprehending larger pieces of connected text. Skill breakdowns at the word level are most often indicative of decoding skill deficits. Decoding proficiency is predictive of reading difficulties (Bell, McCallum & Cox, 2003; Carver, 1998; Christo & Davis, 2008; Ehri & Wilce, 1987; Fletcher, Lyon, Fuchs & Barnes, 2007; Good, Simmons & Kame-euni, 2001; Weisner, 2012;). Therefore, it is essential that decoding skills be intact to promote overall reading development. As noted in the multiple theories of reading described earlier, decoding is a necessary component for reading, but in isolation it is not sufficient for reading. Accurate and efficient decoding skills free up cognitive resources for readers to expand vocabulary, understand new material, and make connections between concepts that they read. Decoding allows readers to identify the words that they can then connect meaning to. As decoding skills develop they are refined and expanded through instruction in the elementary grades. This pattern is also reflected in the organization of decoding skills in the instructional standards used in schools

Decoding is included in the instructional standards used in schools. The K-5 English and Language Arts Common Core State Standards (CCSS) includes decoding skills under the “Phonics and Word Recognition” category of the Reading Standards (National Governors Association, 2010). Decoding in the CCSS follows the developmental trajectory that research on the development of decoding supports. That is, decoding skills typically follow a progressive developmental pattern that begins with mastery of simple skills, such as 1:1 letter-sound associations. In the CCSS, those skills are expected to be mastered in kindergarten. Mastery of those associations are critical, as students in grades 1-2 are expected to decode one and two syllable words, and recognize and apply decoding rules for additional letter patterns (e.g., blends, digraphs, trigraphs, and vowel teams). Mastery and efficiency in identifying more complicated letter patterns is then expanded further in grades 3-5, with expectations that students recognize affixes, root words, and decode multisyllabic words with variant spellings. The standards for each year are based on the assumption of mastery of the previous years standards. Students that experience difficulty with basic decoding skills in the primary grades are unlikely to master the more complex skills in later grades.

Assessment

As students master skills in reading across grade levels those students that experience skill delays or deficits are at a disadvantage both in the grade level that the skill breakdown occurs in, and, without remediation, in subsequent grades. As proficient students continue to advance over time, struggling students continue to fall further behind (Stanovich, 1991). Students that struggle to read in early grades, rarely catch up to their non-struggling peers without appropriate intervention (Torgesen, 1998). This Matthew Effect, or the widening of the gap between students who are proficient and students who struggle over time, must be remedied via accurate identification processes and efficient and effective interventions. Using assessment data effectively is one possible way to combat the Matthew Effect, as test data can be used to inform instruction. This makes it more likely that the instruction will address specific instructional needs. Assessments can serve a variety of purposes in instructional decision-making, and the purpose of the assessment should drive the selection of any particular test. In addition to selecting a test that meets a specific purpose, there are several features of the test that should be considered during test selection.

Test Reliability

It is critical that tests used for instructional decision-making have adequate psychometric properties. Adequate reliability and validity are characteristics that teachers should evaluate when selecting a test. Tests with inadequate reliability or validity can lead to misidentification, misdiagnosis, or misinformed decision-making (Rathvon, 2004). Reliability refers to the consistency of a test result. It accounts for error in testing and increases the likelihood that the result represents the true performance of the test taker. Reliability is traditionally reported as a coefficient with values between 0.00 and +1.00. These coefficients are typically calculated using item response theory (IRT) functions, or other analyses of error ratios. The Standards for Educational and Psychological Testing recommend that every test score be supported with evidence of reliability or precision (American Educational Research Association (AERA), 2014). There are a variety of methods available to determine the reliability of a test. Each method (e.g., test-retest reliability, alternate form reliability, split-half reliability, or measures of internal consistency) for reporting reliability results in reporting of a reliability coefficient. The type of evidence of reliability reported is dependent on the test design and purpose. The Standards for Educational and Psychological testing require that information on every reliability calculation be detailed and reported so test consumers are aware of the sample and method used for determining reliability (AERA, 2014). While ideal reliability coefficients approach 1.00, the standards for acceptable reliability coefficients vary based on the purpose of the test (Oosterhof, 2003). For screening purposes, coefficients of .8 or greater are acceptable, while for diagnostic purposes, acceptable coefficients must be at least .9. Tests with higher stakes (e.g., eligibility or placement tests) have higher reliability coefficient requirements (Salvia & Ysseldyke, 2001).

Test Validity

Reliability is a necessary condition for valid measurement (Salvia, Ysseldyke & Bolt, 2013). Examining the evidence of validity is critical when selecting a test. Evidence of validity refers to how well the test measures what it claims to be measuring. Unlike reliability that is directly measured, validity is inferred based on the type of evidence provided. According to the Standards for Educational and Psychological Testing, validation is based on the accumulation of evidence that supports score interpretations (American Educational Research Association, 2014). Evidence of test validity can be presented in multiple ways. The more detail that is provided supporting a stated use of a test the stronger the evidence of validity is and the more confidence a test consumer has in the results of the test meeting its stated purpose. Evidence of validity is typically provided in multiple ways. Evidence of criterion-validity consists of comparing the results of the test of interest to results of other tests that report to measure the same or similar skills to see if the test of interest aligns (i.e, provides evidence of concurrent validity) or predicts (i.e., provides evidence of predictive validity) the results of other tests. Evidence of criterion-validity is of particular interest to teachers who use scores on classroom assessments to predict how students may preform on future tests. Evidence of content-related validity would involve ensuring that the content of the decoding test aligns with the findings of current decoding research. To provide evidence for content-related validity, content area experts may review a test and affirm that the content of the test is representative of the content domain of interest. Test publishers typically include a discussion of the opinion of expert reviewers as evidence of content validity. In addition, detailed reports describing and classifying test items serves as evidence for alignment of the test with the content it contains. Evidence of construct-related validity should be provided by comparing the test of interest to other tests that are accepted as measures of the same construct. Evidence of construct validity is usually reported as multiple pieces of evidence that converge to support that the construct of interest is captured. Examples of construct validity evidence include: factor analyses results, correlation coefficients, and ANOVA results.

The Function of Nonsense Words in Testing

Decoding tests typically require students to read lists of words. However, it is important to control for memory when assessing decoding skills. Tests that use real words to assess decoding run the risk of inflated scores as students may recognize words by sight, and therefore do not rely on their decoding skills to pronounce the word. To counter this, decoding tests may be composed of nonsense words. Using nonsense words as the mode of gauging decoding abilities requires readers to apply their grapheme-phoneme correspondences to read the word, and eliminates the possibility that the reader is able to read the word based on memory. Students who are able to fluently read nonsense words have internalized the codes found in printed text with enough efficiency that they can apply them to reading nonsense words (Gough, 1996). Reading nonsense words reflects familiarity with orthographic patterns and can indicate the range of a reader’s knowledge of word patterns (Pierce, Katzir, Wolf & Noam, 2010). Proficient and struggling readers will display differences when reading nonsense words. Struggling readers will display impairments when reading nonsense words compared to the performance of their proficient peers (Gottardo, Chiappe, Siegel, & Stanovich, 1999). The ability to read nonsense words is predictive of the ability to read real words and of overall reading level (Carver, 2003). This makes nonsense word reading the preferred mode of assessing decoding skills. For students who struggle in reading using nonsense words to measure their decoding abilities is an efficient method for identifying decoding skill deficits.

Decoding Fluency

An additional consideration in decoding tests is accounting for fluency of skills. Students who are able to quickly and accurately apply decoding rules to read words in isolation (e.g., word lists) are more likely to be proficient with applying skills in connected text (Ehri, 1980). Decoding tests should provide some indication of skill automaticity. However, under timed constraints (e.g., one-minute timed measures) readers will have limited opportunities to demonstrate skills. This is particularly problematic for readers who struggle, and take more time to read words. Timed fluency tests for decoding limit the number of words the student is able to read and decreases the number of patterns or skills that the teacher may observe during the testing session (Ritchey, 2008; Reutzel, Brandt, Fawson, & Jones, 2014). However, decoding fluency is still an important consideration because laborious application of grapheme-phoneme correspondences does not indicate proficiency (Joshi & Aaron, 2002). As an alternative, fluency can be calculated by using the total amount of time that a student takes to read a complete set of words and then dividing the number of words by the number of minutes it took to complete the entire task. This method of measuring fluency does not limit the number of test items a student reads, but could still be a time consuming aspect of testing, particularly for students who struggle with decoding. Decoding tests should strike a balance between the number of skill demonstrations offered and the time it takes to complete the assessment task.

Decoding Assessments Over Time

The Mental Measurements Yearbook (MMY) is published every 2-5 years by the Buros Center for Testing at the University of Nebraska-Lincoln. The first volume of the MMY was published in 1938. Since, then there have been 19 additional published volumes, with the latest published in 2016. The MMY is meant to communicate pertinent test information to a consumer audience to promote informed test selection. Typical MMY reviews include descriptive test information, as well as two test reviews completed by independent professionals from both academic and practical fields. According to the Buros Center for Testing in order to be considered for review by the MMY test products must be “commercially available, be published in the English language, and be new, revised, or widely used since last appearing in the Mental Measurements Yearbook” (Mental Measurements Yearbook, para. 2). In addition, beginning with the fourteenth volume tests must provide some evidence of technical adequacy to be considered for inclusion.

Decoding tests have been included in all volumes of the MMY. The minimum number of decoding tests included any MMY volume was two (MMY Vol. 2, 1940; MMY Vol. 11, 1992). The maximum number of decoding tests included in any MMY volume was 27 (MMY Vol. 8, 1978). Over time, the number of decoding tests included in the MMY has fluctuated. This is likely due to influences of educational trends and policies. Increases in the number of decoding tests in the late 1950s coincide with increases in demand for phonics instruction in classrooms that was the result of response to Robert Flesch's publication of "Why Johnny Can't Read – And What You Can Do About It" in 1955. The largest peak in number of decoding tests in the MMY (i.e., Vol. 8 in 1978) follows the passing of Public Law 94-142 in 1975. With decoding being the most likely source of reading difficulties, and most students with disabilities experiencing some difficulty with reading, it follows that the number of decoding tests increased as students with disabilities were included in school settings. The large decrease in the number of decoding tests included in the MMY in the late 1980s and early 1990s is associated with the rise in popularity of whole-language classroom instruction which de-emphasized phonics components of literacy instruction (Pearson, 2014). The number of decoding tests began to rise again in 2001, likely in response to the passing of No Child Left Behind (NCLB) and the testing and accountability requirements it mandated. Volume 20 of the MMY (i.e., the most recent publication) included five decoding test reviews. 182 decoding tests are included across the twenty volumes of the MMY.

In addition to fluctuations in the number of decoding test reviews included in the MMY there have also been changes in the information and characteristics of the tests included in the MMY since 1938. The number of tests that report psychometric properties has steadily caught up to the total number of tests included in the MMY. In the first ten volumes of the MMY only 49% of included decoding test reviews included any reliability information, and only 30% included any validity information. In volumes 11-20, 86% of included decoding test reviews included reliability information, and 78%

included validity information. However, the number of tests that report to use nonsense words as the mode for testing decoding remains low. Nonsense words were first noted in decoding test reviews in the 1959 publication of the fifth volume of the MMY, with one review that noted nonsense words were used on the measure of decoding. Since then, there was a slight increase in the seventh (published in 1972) and eighth (published in 1978) volumes of the MMY with up to 55% of the included test reviews in those volumes reporting nonsense words were used on included decoding tests. However, the percentage of decoding tests in the most recent volumes of the MMY are lower with only 20% of decoding tests reviewed in volume 20 (2016), 33% of the decoding tests reviewed in volume 19 (2014), and 22% of the decoding tests reviewed in volume 18 (2010) report that nonsense words were used on the included tests. In total, only 26% of the decoding tests included in the twenty volumes of the MMY report using nonsense words.

General Role of Assessments in Public Schools

The role of data and assessment in schools in the United States has evolved over time. In the 1960s school districts began to adopt and administer norm-referenced tests as a means of assessing student achievement. The role of large scale testing expanded to statewide assessment practices in the 1970s. Coinciding with the increase in prevalence of standardized testing there was movement at the national level to develop nationwide testing programs to assess the general success of students across the United States (Stiggins, 2008). The role of assessment also expanded to the international level in the 1980s when comparisons between student achievement scores in the United States were compared to student scores in other countries. In 1983, the seminal report: *A Nation as Risk* alarmed the population of the United States that students were scoring significantly behind their international peers. This resulted in a rise in attention paid to assessment results for schools as policymakers searched for ways to compare student performance and measure school improvements (Vinoviskas, 1998). The National Assessment for Educational Progress (NAEP) took on new importance as a tool that could be used to gauge the health and success of American schools. In addition, recent educational legislation mandates that assessment data be used for accountability purposes.

The No Child Left Behind (NCLB) Act of 2001, brought widespread attention to the efficacy (or lack of efficacy) of schools in the United States, and ushered in the era of accountability in public schools. NCLB was largely focused on holding schools accountable for the progress and achievement of all students (Yell, 2016). NCLB created accountability requirements that all states and schools were required to adopt. NCLB required that states develop or adopt annual tests that measured student progress in reading, mathematics, and science. All students in grades 3-12 were required to participate in accountability testing with at least 95% of the population of student subgroups participating. Examples of subgroups included: students with disabilities, students with limited English proficiency, socioeconomic groups, and racial and ethnic minorities. Schools were now accountable for the proficiency of all students regardless of subgroup. The oversight for schools was held at the federal level. Over time the percentage of overall students, and percentage of students in all subgroups meeting proficiency standards would increase with goals of 100% of students meeting proficiency standards by 2014. Increasing the percentage of students meeting proficiency standards over time was a means of closing the achievement gap between various groups of students (Linn, Baker & Betebenner, 2002). Despite the accountability mandates, NCLB failed to meet its 100% proficiency goal in 2014. However, its failure opened the door to new educational policies that are attempting to promote reading success and address the remaining concerns over the efficacy of public schools in the United States.

The Every Student Succeeds Act (ESSA), passed in December 2015, maintains the same testing requirements mandated by NCLB, but returns oversight to the states from the federal level. The ESSA also introduces language that states can adopt that can cap the amount of time that students spend taking assessments each year, and states can choose to allow schools to use a variety of assessments as indicators of progress, and not just statewide standardized tests. This change shifts the focus of public education in the United States from testing to instruction, but continues to maintain that assessment does have a role instructional practice.

Test Format

The format, or visual layout, of decoding tests has remained largely unchanged over time. The traditional format of decoding tests involved asking students to identify words in isolation, usually formatted as individual words in rows or columns. The first decoding tests were primarily designed to be receptive measures. On those tests, students were asked to silently read a word and to select a definition or picture that matched the word. This method allowed for decoding to be assessed in a group setting. However, over time more tests began requiring students to read aloud to demonstrate decoding skills. Transitioning from silent reading to reading aloud allowed teachers to capture and record the specific error that the student was making when reading. Teachers that could document specific errors or error patterns could more easily identify specific skills that the student was struggling with and provide additional instruction to remediate those skills. On most decoding tests students were asked to identify words presented either one at a time, or in word lists. This format remains the typical format of current decoding tests. Curriculum-based measures (CBM) of nonsense word fluency (NWF) require students to read lists of words on a single page of paper, laid out in horizontal rows (e.g., DIBELS, Fast Bridge Learning, Aimsweb). More formal measures of decoding, such as the Word-Attack subtest on the Woodcock Reading Mastery Test (WRMT), or Phonemic Decoding Efficiency section on the Test of Word Reading Efficiency (TOWRE) require students to read lists of words in single vertical columns. To date, there has been little research on the effect of test format on student performance. The selection of test format appears to be at the digression of the test developer, and once a format is selected subsequent versions of the test retain the original format. The limited research on format for reading tasks is detailed below.

Typography

The typography used in the printed text influences reading performance of young children. In their 2009 studies, Wilkins, Cleave, Grayson, and Wilson found that elementary students (ages 7-9) reading performance was sensitive to changes in the typeface used in reading material. The authors found that younger students read 9% faster when the typeface is larger (26 pt. font) versus smaller (22 pt. font). In addition, the authors noted that many reading tests are formatted that as word difficulty increases, the font size of the presented text decreases. Results show that there is an increase in performance when the font size remains constant across the entirety of a reading task, suggesting that young readers are sensitive to changes in text size. Last, Wilson et al., found that the font selection impacts the reading performance of young readers. In their sample, participants read faster and with more accuracy when the font used matched common reading text typography than when a juvenile style font (i.e., fonts with stick and ball style or fonts with the single story ‘g’ instead of double-story ‘g’, etc.) that matched the common writing style of the participants. Font selection and style are likely to impact the reading performance of young readers, and should be given consideration when formatting reading tasks for this population.

List Reading

Reading words in lists has become the typical presentation of decoding tasks. Research studies have confirmed that list reading (i.e., reading individual words that are not connected by context) provides an adequate indication of reading proficiency (Ehri, 1980; Jenkins, Fuchs, Van den Broek, Espin & Deno, 2003). In Linnea Ehri’s 1980 study on beginning reading she examined differences in the reading performance of students who read sentences (i.e., sentence readers) versus students who read words in isolation (i.e., list readers). Results show that list readers were able to remember more of the orthographic features of words, and were therefore able to generalize reading the word to other contexts, and more accurately spell the words. Reading words in isolation forced the reader to rely on decoding skills to pronounce unknown words, while sentence readers could rely on context to correctly guess unknown words. The disadvantage of sentence reading as a measure of student reading performance, Ehri noted, was that context cues contribute to correct guessing which results in a decrease in opportunities for a teacher to provide corrective feedback to promote word reading skills. List reading controls for semantic cues and provides a cleaner and more accurate indication of students decoding abilities. However, if real words are used in list format there are still variables, such as memory, that may limit how indicative results are of a readers decoding skill. Therefore, to measure decoding nonsense words are still the preferred method.

In their 2003 study of the accuracy and fluency reading, Jenkins, Fuchs, Van den Broek, Espin and Deno found that list reading was a sensitive mode for detecting reading impairments. Their study compared the performance of skilled and impaired readers contextual reading (i.e., reading a paragraph of connected text) to list reading (i.e., reading a randomized list of the words included in the connected text passage). Their results show that skilled readers outperformed impaired readers regardless of text format. All participants that displayed impairment in context reading also displayed impairment in list reading (i.e., the list reading task was able to discriminate the impaired readers from the skilled readers). Impaired readers are significantly discrepant from their skilled peers regardless of the format that they were asked to read.

In follow up analyses of the Jenkins, et al., study the authors further examined the utility of list, or context-free reading tasks. In 2003, Jenkins et al., published additional analyses that explored the role of context free reading as it connected to context fluency. The authors found that readers who are able to fluently read words in isolation are more likely to read fluently when reading pieces of connected text. List fluency, they concluded, is a pure measure of word-reading efficiency, as it controls for individual differences in phonological awareness, orthographic memory, and naming speed. Measuring list reading fluency can provide an indicator of broader reading proficiency.

Number of Items

The number and organization of words presented to a reader at a time has an effect on reading performance as well. In their 2009 study of the effects of various matrix presentations for a letter identification task, Jones, Branigan, and Kelly found that identification performance depended on the skill of the reader. Their sample included a group of participants diagnosed with Dyslexia. The authors compared test performance of the group with Dyslexia to the performance of a group without Dyslexia. Results show that the group with Dyslexia displayed impairments in identification when multiple items were presented simultaneously (e.g., 50 letters spaced evenly on a single page). In contrast, the performance of the group without Dyslexia was enhanced when multiple items were presented simultaneously versus individually (e.g., one letter per page). The findings support the idea that proficient readers benefit from multiple items because they are able to efficiently process information quickly, and benefit from preview of subsequent items. It is likely that impaired readers do not experience benefit from previewing subsequent items on a test. This hypothesis has been supported by additional studies of eye tracking and preview benefit.

In studying the eye movements of readers researchers have found that there is a connection between reading skill and preview benefit. Preview benefit is the increase in speed of identification of the word following the word a reader is currently fixated on. In English texts, preview benefit refers to the word to the immediate right of the word a reader is fixated on. In 2005, Chace, Rayner, and Well, found that skilled readers benefit from preview of the next word in reading, while unskilled readers display no preview benefit. The researchers hypothesized that this was likely related to unskilled reader requiring full attention to the fixated word, and therefore having few cognitive resources left to obtain information from upcoming words. This finding supports the theory of automaticity that unskilled readers devote full cognitive resources to each individual word, leaving few resources available for other processes (Perfetti, 1986).

Additional studies by Frömer, Dimigen, Niefiend, Krause, Kleigl and Sommer in 2015 and Marx, Hutzler, Shuster and Hawelka in 2016 support the connection between preview benefit and reading skill. Frömer et al., found positive correlations between reading fluency and preview benefit. They concluded that more fluent and skilled readers take in more information from the parafoveal word (i.e., the next word in a sequence) than less skilled readers. Marx et al., extended the eye tracking and preview benefit research to young readers, and found results that support that even readers in the beginning stages of reading development demonstrate preview benefit. Their results show that reading competency is the best predictor of preview benefit. It accounts for more of the variance in preview benefit than age, grade, or reading experience. In addition, Marx et al., found that the readers with the greatest decoding skills exhibited the greatest preview benefit.

Student Motivation and Preference

An additional consideration in reading performance is student motivation for reading, and their preference for presentation of items. There has been extensive research on the relation between motivation to read and reading performance. In their meta-analyses of reading motivation research, Morgan and Fuchs (2007) found that the research on motivation and reading performance supports a bidirectional relation between motivation to read, and reading performance. That is, students who are motivated to read are likely to experience more success with reading, and students who experience success when reading are likely to be more motivated to read (Morgan & Fuchs, 2007; De Naeghel, Van Keer, Vansteenkiste & Rosseel, 2012). The correlation between reading skill and reading motivation shows that a potential avenue of addressing low reading skill may be in addressing low motivation. The research on motivation and reading skill also references the problems with the Matthew Effect. Students who experience both reading difficulty and lowered motivation in reading are likely to fall further behind their proficient peers over time. While, it is impossible to determine a causal relation between low reading motivation and poor reading performance, Morgan and Fuchs, do advise that addressing both pieces of the interaction (i.e., motivation and skill) as early as possible may prevent future reading failure and counter the Matthew Effect that poor readers experience over time. Additional studies have shown that improving reading performance results in increased motivation for reading. In their 2014 study, Chen and Savage found that providing phonics instruction to struggling readers resulted in improvements in both reading performance and reading motivation.

With motivation identified as a pertinent factor to explore in reading performance. It is important to examine the role it may play in testing situations. One aspect of motivation, the role of student choice, can directly impact reading performance. Students who are given choice in reading materials are likely to experience increases in motivation, put forth more effort in the learning task, and experience increases in performance (Randi & Corno, 2000; Ciampa, 2016). While there is little research that has examined the role of choice in testing, there are logical connections from the research on the influence of student choice and preference that might be applied to testing. When students are given the power to choose, and control elements in their learning environment they are more likely to engage in their tasks (Beishuizen, 2008).

While no studies have been done that have examined student preference for testing formats in elementary students, there is some, yet limited, research on testing preferences in middle school students. Student preferences for testing adaptations may influence the effectiveness of those adaptations at promoting student performance (Polloway, Bursuck, Jayanthi, Epstein & Nelson, 1996). Using a survey, Nelson, Jayanthi, Epstein and Bursuck (2000) investigated the testing adaptation preferences of middle school students. Their results show that the highest preferred adaptations were the least obtrusive (i.e., involved no verbal interaction, and were not noticeably different then the standard test format). Nelson et al. also found that students that experience low achievement had higher preferences for testing adaptions than did students with higher achievement. The authors hypothesize that struggling students recognize the role of testing in the classroom, and are self-aware of test adaptions that may enhance their performance.

Computer Platform for Testing

An additional facet of test format investigation is the consideration of the role of technology platforms for test administration. The ratio of available computers to students is 1.6:1 in elementary schools, with 61% of teachers reporting that they use technology to administer assessments (Gray, Thomas & Lewis, 2010). Computer based tests (CBTs) are becoming increasingly prevalent in test publications. CBTs are more prevalent in large-scale assessment administrations (e.g., National Assessment for Educational Progress (NAEP, NCES, 2015); Partnership for Assessment of Readiness for College and Careers

(PARCC, 2014); Smarter Balanced Assessments (SBAC, 2014)). However, there has been an increase in the number of formative classroom assessments that are making use of available technology as a means of increasing efficiency and effectiveness (Brown, Hintz, & Pellegrino, 2008). Advantages of CBTs include: instant grading and score tracking, fewer materials to store and purchase, guaranteed standardized presentation of materials, decreased administrator error, immediate feedback, and additional data collection (e.g., latency scores) that are otherwise not collected in a paper administration of a test (Bodmann & Robinson, 2004; Quellmalz & Pellegrino, 2009). Several CBM publishers have begun to incorporate technology into their products. The Dynamic Indicators for Basic Early Literacy (DIBELS) Next system has a computerized version, called mClass, for scoring and data analysis. FastBridge Learning and Aimsweb publishers have technology administration options for their early literacy measures. FastBridge Learning, Aimsweb, and EdCheckup publishers have technology administration options for reading CBM measures (Hosp, Hosp, & Howell, 2016). However, the format of the technology based CBM administrations does not differ from the paper version of the tests. Instead computer versions of CBM instruments simply mimic the paper format of the test by presenting the same test items, in the same layout (e.g., rows of words) on the screen.

It is important to consider several factors when moving from a paper-based test to a CBT. In their 2006 study comparing performance of paper and CBT reading comprehension test performance in primary grades, Pomplun, Ritchie, and Custer found that student characteristics, such as computer familiarity, contributed more to score differences between the two modes than did item characteristics (i.e., how difficult an item was). It is likely that as students are exposed to more technology throughout their time in school, and the prevalence of technology continues to increase, that the effects of computer familiarity on test performance will decrease.

CBTs have been found to be a more efficient method for collecting a variety of data across a variety of populations. Carson, Gillon, and Boustead (2013) found that CBT administration reduced test time by 20% on a measure of phonological awareness in primary students when compared to the administration time of the paper-based measure. The decrease in testing time resulted in no difference in the reliability of the score. Bodmann and Robinson (2004) found that undergraduate students completed CBTs significantly faster than paper multiple-choice tests while producing equivalent scores. To date, no studies have examined the characteristics of CBTs for decoding skills.

The Importance of Format in Test Taking

Given that studies have shown that using CBTs is an option for improving test efficiency it is also necessary to examine the format of the CBTs to determine if there are ways to improve the efficiency of test administration even more while maintaining the integrity and accuracy of the test. In addition, accounting for student preference in test format selection, and incorporating information from research on list reading may lead to test layout improvements that benefit the test taker, by optimizing performance, and the teacher, by decreasing the time spent testing.

One study has looked exclusively at the effects of format on student performance on a classroom assessment. In their 2015 exploratory study, Jones, Gifford, Yovanoff, Al Otaiba, Levy, and Allor examined how alternate assessment formats for progress monitoring sight word reading impacted the performance of participants with intellectual disabilities. In their study they tested three format conditions: PowerPoint presentation of individual words, flashcards with individual words, and a paper and pencil format (i.e., original format of test). Their results show that changing the format of the test task increased the reliability of student performance without compromising the test accuracy, in a sample with participants that traditionally struggle with test engagement. The results of this preliminary study support the hypothesis that altering test format may affect test performance. It is critical that tests are designed to accurately and efficiently measure student performance so that teachers can quickly use the data to determine what instruction the student may need. Improving efficiency in testing by altering test format may impact the instructional quality that the student receives as a result of their performance on the test.

Purpose of the Study

The purpose of the current study is to examine the effects of test format on student decoding performance. With decoding playing a critical role in reading development, tests that measure decoding must be efficient and effective at informing instructional decisions in the classroom. As nonsense word reading is the preferred method of assessing decoding skills, all format tasks in this study will include equivalent nonsense words. In this study participants will engage in decoding tasks (i.e., reading nonsense words) arranged in a variety of formats on a computer. The accuracy and test time across all the formats will be measured and compared to determine if an optimal test format for measuring decoding exists. In addition, student preference for format will be solicited from participants to examine the connection between student choice and performance across formats.

Research Questions

This study seeks to answer the following research questions:

1. Does test format affect decoding accuracy and fluency in second grade students?
2. Does student preference for test format affect decoding accuracy and fluency in second grade students?
3. How do decoding tasks, presented in different formats, relate to other measures of reading?

CHAPTER 3

METHODS

Overview

The purpose of this study is to examine the effects of test format on student decoding performance. To answer the three research questions a variety of techniques were used. This chapter will provide information on the study design, setting, participants, and the procedures used to answer each of the research questions. In addition, the instruments used to collect data for analyses are described. This chapter begins with providing general study information on the design, setting, participants, and procedure of the study. Then the chapter is split into sections describing the methods and analyses that were used to address each research question.

Design of this Study

To answer the first research question regarding student performance differences on different test formats a repeated measures design was used. This design allowed for multiple comparisons within the same participant. In this study, each participant completed a decoding task presented in five test format conditions. The five conditions were: 5 word columns, 5 word rows, 2 word columns, 2 word rows, and single word presentation. The results on each condition were compared to the scores on other conditions to determine performance differences across test format conditions. These data were also used to answer the second research question.

To answer the second research question regarding the relation of student preference for format to performance, student preference selection was gathered from participants. Performance on the selected preferred format was compared to student performance on the other (i.e., non-preferred) formats. Differences in performance related to student choice provided data regarding the influence of student preference for format on decoding test performance.

To answer the third research question regarding how performance on different test format conditions relates to performance on other decoding and reading tests, comparisons between several dependent variables were completed. Additional dependent variables included: DIBELS Next nonsense word fluency (NWF) scores, DIBELS Next oral reading fluency (ORF) scores, the Decoding Inventory for Instructional Planning Screener (DIIP-S) scores, and scores on a broad measure of reading ability (i.e., the Group Reading and Diagnostic Evaluation (GRADE). Comparing scores on the test format conditions to performance on several dependent variables provided evidence of criterion-related validity of the decoding task given in different formats.

To determine the number of participants necessary for this study a power analysis was conducted using R statistical computing software (The R Foundation for Statistical Computing, 2013). The power analysis to calculate sample size included power set at .8, alpha set at .05, and the selected effect size (Cohen's *d*) of .5. The results of the power analysis show that the minimum sample required for this study was 34 participants.

Setting and Participants

Participants in this study were students in Grade 2 (n=53) from a small city in the northeast. Participants attended an elementary school that served students in grades 2-4. The elementary school is one of four schools that comprise the school district. This study was conducted with second graders because decoding skills are emphasized in second-grade reading standards, and typically are intact by the end of the second grade year

(Chall, 1996). By selecting second grade students for this study it also increased the likelihood that the sample would include a large distribution of decoding skills. Some students would have mastered basic decoding skills, while others would still be learning letter-sound correspondences. In addition, typical reading instruction during second grade transitions from individual decoding skill instruction to fluency and application of decoding skills. Students who have not mastered basic decoding skills, like those measured in this study, are likely to experience difficulty with other reading skills (Ehri, 1998).

Participants were recruited from three second-grade classrooms that had volunteered to participate in this study. All students in these classrooms were eligible for participation. Parental consent forms were sent home to 62 students. A total of 54 consent forms were returned. During administration, one test session was disrupted, and that participant was removed from the study. Demographic information for the 53 participants can be found in Table 1.

Table 1: Demographic Participant Data (n=53)

Group	Percentage of Participants
Race	
White	83.0
African American	3.8
Asian	1.8
Latino/Hispanic	5.7
Multi-race, Non-Hispanic	5.7
Gender	
Male	43.4
Students with Disabilities	1.8
English Language Learners	5.7
Students Receiving Title One Reading Services	26.4

Note. The percentage of students that qualified for free or reduced lunch was not available for this sample. However, the percentage of students that qualified for free or reduced lunch across all three participating classrooms that the sample was pulled from was 34.9%.

Study Procedure

In this study, each participant participated in two testing sessions: one group session to complete the GRADE, and one individual session to complete the Dynamic Indicators of Basic Early Literacy (DIBELS) Next NWF measure, DIBELS Next ORF measures, the Decoding Inventory for Instructional Planning-Screener (DIIP-S), and the 5-condition test format task.

Instruments

Participants in this study completed multiple reading tasks. For all published tests standardized directions and scoring procedures were used.

The DIBELS Next NWF measure that was used in this study is an individually administered and timed measure of a student's ability to apply letter-sound correspondences to pronounce nonsense words (Good & Kaminski, 2011). Students are given one minute to read nonsense words that are presented on a single sheet of paper. At the onset of the NWF task students are guided through two practice nonsense words to orientate them to the task. They are provided corrective feedback during the practice words. Following the practice words students are given the following directions for the actual task: "I would like you to read more make-believe words. Do your best reading. If you can't read the whole word, tell me any sounds you know. Put your finger under the first word. Ready, begin." At the end of one-minute the data collector marked the last word or sound the student had pronounced and the student was asked to stop. The nonsense words follow two patterns: CVC and VC. In this study students completed the fall second grade benchmark measure. DIBELS Next NWF results in two scores: Correct Letter Sounds (CLS) and Whole Words Read (WWR). CLS scores are based on the total number of correct letter-sound correspondences the student produced (i.e., the total number of correct phonemes pronounced in one minute), while WWR scores require students to correctly pronounce the nonsense word as a complete syllable without segmenting any individual phonemes (i.e., WWR scores are the total number of complete nonsense words that are pronounced in one minute). For example, students who read the nonsense word "bav" by producing the sounds /b//a//v/ individually received credit for 3

CLS, and a score of 0 for WWR. By comparison, students who read “bav” by producing /bav/ (i.e., with all correct phonemes correctly blended) received credit for 3 CLS as well as a score of 1 for WWR. It is possible for students to have received partial CLS credit. For example, a student who produces the /n/ and /p/ for the word “nop” received a score of 2 CLS, and 0 for WWR. Both CLS and WWR scores were used for analyses. Alternate-form reliability for DIBELS Next NWF is .90-.96 for WWR, and .85-.94 for CLS. Test-retest reliability is .70-.88 for WWR, and .76-.90 for CLS. Interrater reliability is .99-1.00.

The DIBELS Next ORF measures that were used in this study are individually administered, and timed measures of a student’s reading fluency (Good & Kaminski, 2011). In this study, participants completed the fall second grade benchmark ORF task. Participants were read the following directions at the onset of the task: “I would like you to read a story to me. Please do your best reading. If you do not know a word, I will read the word for you. Keep reading until I stay “stop”. Put your finger under the first word. Ready, begin.” To complete the ORF benchmark measures students were given one minute to read aloud from a grade level passage. The benchmark procedure has students read three ORF passages, one right after the other, and they are each presented on a single sheet of paper. The median scores on the three passages was used to determine if the student met benchmark goals. In addition, the median score was used for analysis in this study. The benchmark resulted in three words read correctly (WRC) scores, one for each passage. WRC scores are the total number of words that the student correctly identified in the one-minute reading sample (i.e., this is the reading fluency score). Errors are considered any word that is substituted (e.g., /talk/ for /took/), or omitted (e.g., reading "the cat ran quickly" for "the black cat ran quickly"). In order to receive credit each word in the text must be pronounced as it is written. The median WRC score was used in analysis. Alternate-form reliability for ORF passages is .89-.96 for WRC. Test-retest reliability is .91 for WRC. Interrater reliability is .99.

The DIIP-S measure that was used in this study is an individually-administered test of a reader's decoding skills (Hosp, 2016). It is based on the Decoding Inventory for Instructional Planning: Diagnostic (DIIP-D); a diagnostic test of decoding that includes around 600 words. To complete the DIIP-S students are required to read 33 nonsense words and 3 contractions, for a total of 36 words, representing 12 decoding skills (i.e., 3 words per skill). The items included on the DIIP-S were selected from the DIIP-D as the three words with the best discrimination parameters within each of the 12 decoding categories. The nonsense words on the DIIP-S are representative of the 11 most common decoding patterns in English. Each included decoding skill and an example of a word representing that skill are presented in Table 2. The directions for the DIIP-S were provided at the beginning of the task and were: "I want you to read some words to me. These are not real words, except for the last three. The rest are all made-up words. I want you to try your best to read each word. Point to each word as you read it. Start with the first word here and read across the page. Be sure to do your best reading." The DIIP-S results in two scores: total words read correctly (WRC) and a fluency score of (WRC/minute). To receive credit for a word, the reader must correctly pronounce each word in its entirety. Students who produce each phoneme in isolation must correctly blend the phonemes to receive credit (i.e., words are scored at the whole word level). Unlike the scoring on the NWF measure, no partial credit was given on the DIIP-S. To determine fluency on the DIIP-S the total number of words the student read correctly is divided by the total time it takes the student to complete the DIIP-S. The WRC and fluency scores were used for analyses. Internal consistency on the DIIP-D subtests (i.e., each separate decoding skill) ranges from .857-.963. Test-retest reliability of the DIIP-D ranges from .748-.951 (Robbins, Hosp, Hosp, & Flynn, 2010).

Table 2: Decoding Skills and Example Nonsense Words on the DIIP-S

Decoding Skill	Example Nonsense Word
CVC	vod
CVCC	wunk
CVCe	wame
r-controlled vowel	ker
Blend	sneb
Digraph	thod
CVCCVC	gogset
Vowel Team	haid
Prefix	precred
Short Vowel Suffix	magness
Long Vowel Suffix	gotion

Note. C=Consonant; V=Vowel

Participants also completed a modified NWF task. I created this task by formatting a 5-condition, 100-word nonsense word reading task using Microsoft PowerPoint. Each format condition included 20 nonsense words. The words included in the test format task were selected from progress monitoring forms 1-3 of the DIBELS

Next NWF second grade forms. Because progress monitoring forms are designed to be equivalent (alternate form reliability for DIBELS Next NWF is reported as .90-.96 for whole words read) words were selected and randomly assigned to each of the test format conditions. The words included on DIBELS Next NWF forms follow either consonant-vowel-consonant (i.e., CVC) patterns or vowel-consonant (i.e., VC) patterns. To control for word length only CVC words were selected for inclusion on the test format conditions task. To help ensure that each format task was equivalent each vowel (i.e., a,e,i,o, and u) were represented an equal number of times in each condition (i.e., 4 times in each condition). In addition, within each format condition and within each vowel category, no beginning or ending consonants were repeated (e.g., within the 5 word column format there are four CVC nonsense words with “a” as the vowel, and across those four words there are no repetitions of beginning or ending consonants). No words were repeated in the task.

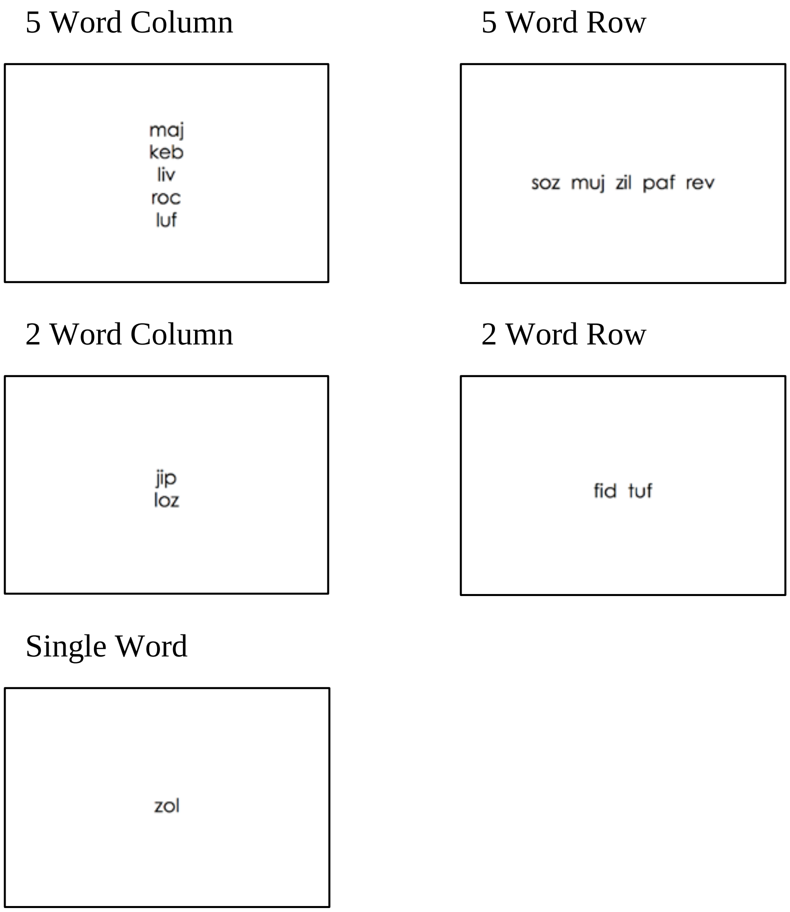
Directions were provided to the student at the onset of the task and were read as follows: “I would like you to read some more make-believe words on the computer. After you read the words on the screen I will move it to a new screen where you will read more words until you are done. Sometimes you will need to read across the screen and sometimes down the screen, and sometimes there will just be one word to read at a time. If you can’t read the whole word, tell me any sounds you know. Ready. Begin.” The data collector controlled the advancement of the slides using the computer mouse. Scoring and administration rules were standardized and based on the DIBELS Next NWF directions. Students received credit for each correct letter-sound correspondence (CLS) that they identify. In addition, students were given credit for each whole word read (WWR).

Students received credit for each WWR by correctly pronouncing the entire word without first producing the individual phonemes.. In the 5-condition test format task both CLS and WWR were calculated for each format. In addition, time was recorded for each format as the researcher noted the time it took for each student to complete each format condition. The time was used to determine a student's fluency (CLS/minute and WWR/minute) on each test format condition. All scores (i.e., CLS, WWR, CLS/minute, and WWR/minute) were used in analyses to determine differences in student performance in each test format condition.

Including twenty words per condition allowed the reader the opportunity to demonstrate enough letter sounds, and total words read, to meet the NWF benchmark for the beginning of second grade in each condition (54 CLS and 13 WWR) (Dynamic Measurement Group, 2010). In each test format condition the student had the opportunity to demonstrate up to 60 CLS and 20 WWR.

The task was presented to the student using Microsoft PowerPoint. The 5-condition test format task included a total of 48 slides. Each slide had a white background with black text. Text was centered on each slide, using Century Gothic, 42-point font. This font matches the font used on the DIBELS Next NWF measures. Columns were single spaced, and there was a double space between each word in row formats. To control for practice effects the presentation order of conditions was counterbalanced. Each participant was given a different order of format conditions. Figures 1 displays examples of each of the format conditions.

Figure 1: Test Format Conditions



5 Word Column

5 Word Row

2 Word Column

2 Word Row

Single Word

The GRADE is a norm-referenced and group-administered, assessment of literacy competency (Williams, 2001). It is designed for use with students from preschool through 12th grade. Participants in this study took the GRADE Level 2 designed for second grade. The GRADE was administered in a group format in each of the participating classrooms. The GRADE consists of four subtests: Word Reading, Word Meaning, Sentence Comprehension, and Passage Comprehension. The Word Reading subtest measures a student's ability to decode or recognize sight words and requires students to select a target word, provided by the examiner, from a field of four words. The Word Meaning subtest measures decoding or sight-word reading, and understanding of early reading vocabulary. It requires the student to silently read a target word and select the picture that matches the word from a field of four pictures. The Sentence Comprehension subtest measures a student's ability to understand a sentence as a whole thought. It requires the student to read a sentence with a missing word and select the correct missing word from a field of four. The Passage Comprehension subtest measures the student's ability to comprehend a variety of content presented in short passages. Students are required to read the passage and answer several multiple-choice questions about the content of the passage. The GRADE results in two composite scores. The vocabulary composite consists of the scores on Word Reading and Word Meaning. The Comprehension composite consists of scores on Sentence Comprehension and Passage Comprehension. The total test score consists of all four subtests. The two composite scores, and the total test score were used for analyses. Internal consistency of the GRADE is .91-.99**.** Test-retest reliability is .77-.96. Alternate-form reliability is .81-.94.

Procedure

Participants completed the individual testing session first. All individual testing sessions were held in the school computer lab. During this session each participant completed student assent, the DIBELS Next NWF benchmark assessment, the DIBELS Next ORF benchmark assessment, the DIIP-S, and the modified NWF 5-condition test format task. In addition to myself, three trained graduate students assisted with data collection. All data collectors completed training that included practice administrations and scoring of each of the measures. Twenty percent of the administrations were checked for reliability using audio recordings of the testing sessions. Reliability was 99.1%. Table 3 indicates the order of the tests, and the estimated administration time for each test for the individual testing session.

Table 3: Test Order and Administration Time

Administration Order	Test or Task	Estimated Administration Time
1	DIBELS Next NWF	2 minutes
2	Test Format Conditions Task	6-8 minutes
3	DIIP-S	3-4 minutes
4	DIBELS Next ORF (3 probes)	4 minutes
Total		15-18 minutes

Note. Estimated administration time includes time for directions. The test format conditions task includes time for collecting preference selection from participant.

The second testing session was completed in each of the participating classrooms. I proctored the administration of the GRADE in each testing session. Teachers assisted in monitoring students, and providing alternate activities for students not participating in the study. The testing session for the GRADE was completed in approximately 30 minutes in each classroom.

Research Question 1

The first research question in this study seeks to determine if test format affects student performance on a decoding task. To answer this question participants completed a modified NWF task on the computer. Four performance scores: CLS, WWR, CLS/min and WWR/min were calculated for each of the five test formats that were included on the task.

To answer research question one, a series of one-way repeated measures ANOVAs were run to determine differences between CLS scores across the five test format conditions, WWR scores across the five test format conditions, CLS/minute fluency scores across the five conditions of the test format task, and WWR/minute fluency scores across the five conditions of the test format task. Results indicate if there are performance differences between the five format conditions on any of the dependent variables.

Research Question 2

The second research question in this study seeks to determine if student preference for test format affects their performance on a decoding test. Following the completion of the 5-condition test format task, described in the previous section, each student was asked to indicate which test format they preferred most. The final slide of the PowerPoint presentation displayed visuals (i.e., thumbnail images) of each of the format conditions. Students were asked to indicate which of the formats they would prefer. Participants were given the following directions: “Good job reading all those words. Now I want to know which of the layouts you liked best. If I was going to have you read more words, which of these would be how you would want them to look?” The thumbnail images of each of the formats were ordered in the same order that the student encountered them during the test. Students were able to either verbally indicate which format they prefer, or point to select the format on the screen. The data collector recorded student preference selection. See Appendix A for an example of the preference indication slide.

To answer research question two, a series of one-way repeated measures ANOVAs was done. Scores (i.e., CLS, WWR, CLS/minute, and WWR/minute) from the five format conditions were organized by student preference. Scores on the students' preferred format were compared to scores on the non-preferred formats. Results indicate if there are differences in student performance attributed to the student preference for format.

Research Question 3

The third research question in this study refers to the relation between student performance on different decoding test formats and other measures of decoding and reading. This question seeks to identify any evidence of criterion-related validity performance on any condition the 5-condition test format task has. To identify evidence of criterion-related validity scores from the 5-condition test format task were compared with student performance on a variety of other measures of decoding and reading. Other measures of decoding and reading study participants completed included: DIBELS Next NWF, DIBELS Next ORF (DORF), the DIIP-S, and the GRADE.

To answer research question three, a series of analyses were conducted. First, a series of correlations (Pearson's *r*) were calculated between results (i.e., CLS, WWR, CLS/min, and WWR/min) of the five test format conditions and results of the dependent variables (i.e., CLS and WWR on NWF, WC on DORF, WRC and fluency on DIIP-S, and the composite scores and total test score on the GRADE). Results indicate how performance on each of the test format conditions relates to performance on other measures of reading and decoding. Then, Meng's z-test for comparing correlated correlation coefficients was done to compare correlations between the various formats and dependent variables (Meng, Rubin, & Rosenthal, 1992). Results comparatively show the relation of the test format condition and other variables. Comparisons allow for interpretation of evidence of criterion-related validity of the test format conditions.

CHAPTER 4

RESULTS

Overview

This chapter details the results of the study on the effects of test format on student performance. First, descriptive statistics are provided for each of the variables included in analysis. Next, results of the series of repeated measures ANOVAs was completed to answer the first research question and determine if test format affected student performance. Then, results of the series of repeated measures ANOVAs completed to answer the second research question regarding the effects of student preference for test format on performance are included. Last, to answer the third research question, correlational data regarding the relation between performance on the five test formats and other criterion reading measures are presented.

Descriptive Statistics

Descriptive statistics for all variables and criterion measures are presented in Table 4. Table 4 presents the mean, standard deviation, minimum, maximum, skewness, and kurtosis values for each dependent variable (i.e., CLS, WWR, CLS/min, and WWR/min) in each of the five test format conditions (i.e., 5 Column (5C), 5 Row (5R), 2 Column (2C), 2 Row (2R), and Single Word (S)), and the descriptive statistics for each of the criterion variables.

Table 4: Descriptive Statistics (n=53)

Variable	M	SD	Minimum	Maximum	Skewness	Kurtosis
CLS						
5C	53.19	5.63	38.0	60.0	-0.87	0.24
5R	51.49	6.03	37.0	59.0	-0.73	-0.32
2C	52.15	5.85	30.0	60.0	-1.56	3.65
2R	52.72	5.61	39.0	60.0	-0.98	0.10
S	52.85	5.22	38.0	59.0	-1.13	0.57
WWR						
5C	12.91	6.20	0	20.0	-0.86	-0.39
5R	11.34	5.78	0	19.0	-0.54	-0.77
2C	12.04	5.75	0	20.0	-0.82	-0.23
2R	12.13	6.29	0	20.0	-0.93	-0.49
S	12.30	6.08	0	19.0	-0.95	-0.47
CLS/min						
5C	67.16	34.40	9.6	165.7	0.73	0.31
5R	63.63	29.31	20.6	139.2	0.73	-0.22
2C	53.82	23.58	8.9	105.0	0.28	-0.88
2R	54.23	22.17	11.7	99.4	0.07	-0.71
S	46.17	17.49	10.9	82.5	-0.03	-0.73

(continued)

Variable	M	SD	Minimum	Maximum	Skewness	Kurtosis
WWR/min						
5C	17.95	13.27	0	51.4	0.55	-0.47
5R	15.32	11.18	0	43.2	0.66	-0.34
2C	13.39	9.13	0	31.9	0.24	-1.03
2R	13.63	9.03	0	30.9	0.06	-0.92
S	12.13	8.71	0	47	1.03	3.42
DIBELS Next:						
NWF						
CLS	58.43	30.29	8.0	134.0	0.78	0.09
WWR	15.45	11.57	0	44.0	0.67	-0.10
DIIP-S						
WRC	15.77	8.39	2.0	32.0	0.10	-0.89
WRC/min	11.83	10.83	0.50	44.3	1.29	1.05
DIBELS Next:						
ORF						
WRC	74.08	34.96	9.0	164.0	0.56	-0.37
GRADE						
Vocabulary Composite	47.60	8.47	21.0	55.0	-1.36	1.67
Comprehension Composite	19.26	11.32	2.0	45.0	0.58	-0.64
Total Test	66.87	18.18	24.0	99.0	-0.17	-0.54

Note. CLS= correct letter sounds; WWR= whole words read; WRC= words read correctly; NWF= Nonsense Word Fluency; ORF= Oral Reading Fluency; DIIP-S= Decoding Inventory for Instructional Planning- Screening; GRADE= Group Reading and Diagnostic Evaluation.

All included variables were examined for adequate skewness and kurtosis. Skewness and kurtosis values falling between +/- 1.00 and 2.00 are considered questionable and values above/below +/- 2.00 are considered problematic (Tabachnick & Fidell, 2012). In general, all variables displayed appropriate values for skewness and kurtosis. High kurtosis values were observed in the 2C condition for CLS, and in the S condition for WWR/min.

The highest mean score for each dependent variable in the test format condition (i.e., CLS, WWR, CLS/min, and WWR/min) occurred in the 5C condition. The lowest mean scores for accuracy (i.e., CLS and WWR) occurred in the 5R condition. The lowest mean scores for fluency (i.e., CLS/min and WWR/min) occurred in the S condition.

Research Question 1

A series of repeated measures ANOVAs were conducted to compare the effect of test format on decoding accuracy and fluency performance in five test format conditions. A repeated measure ANOVA was run for each of the dependent variables.

Correct Letter Sounds

A repeated measures ANOVA showed that for the 53 participants in this sample the differences in CLS between the five test format conditions were statistically significant, $F(4,49) = 2.629$, $p = .046$, partial $\eta 2 =.177$. Based on guidelines proposed by Cohen (1988) this result suggests a large effect size. To determine the differences in CLS between each test format condition pairwise comparisons were completed to identify the significance between each test format condition. The results of the pairwise comparisons are presented in Table 5.

Table 5: Mean Differences in Pairwise Comparisons for CLS (Standard Error in Parentheses)

Test Format Condition	5R	2C	2R	S
5C	1.70 (0.63)*	1.04 (0.45)*	0.47 (0.52)	0.34 (0.47)
5R		-0.66 (0.62)	-1.23 (0.66)	-1.36 (0.58)*
2C			-0.57 (0.55)	-0.70 (0.45)
2R				-0.13 (0.44)

Note. 5C= 5 word column test format; 5R = 5 word row test format; 2C = 2 word column test format; 2R = 2 word row test format; S = single word test format

*$p < .05$

Based on the pairwise comparisons, significant differences were found in CLS performance between the following test format conditions: 5C and 5R, 5C and 2C, and 5R and S. Differences in CLS score in all other pair comparisons were not statistically significant. In the 5C to 5R and 5C to 2C condition comparisons, participants obtained significantly higher CLS scores in the 5C condition than in the 5R or 2C conditions. In the 5R to S comparison, participants obtained significantly higher CLS scores in the S condition than in the 5R condition.

Whole Words Read

A repeated measures ANOVA showed that for the 53 participants in this sample the differences in WWR between the five test format conditions were statistically significant, $F(4,49) = 3.432$, $p = .015$, partial $\eta 2 =.219$. Based on guidelines proposed by Cohen (1988) this result suggests a large effect size. To determine the differences in WWR between each test format condition pairwise comparisons were completed to identify the significance between each test format condition. The results of the pairwise comparisons are presented in Table 6.

Table 6: Mean Differences in Pairwise Comparisons for WWR **(**Standard Error in Parentheses)

Test Format Condition	5R	2C	2R	S
5C	1.57 (0.43)***	0.87 (0.40)*	0.77 (0.33)*	0.60 (0.33)
5R		-0.70 (0.39)	-0.79 (0.40)	-0.96 (0.41)*
2C			-0.09 (0.42)	-0.26 (0.30)
2R				-0.17 (0.38)

Note. 5C= 5 word column test format; 5R = 5 word row test format; 2C = 2 word column test format; 2R = 2 word row test format; S = single word test format.

* *p* <.05; *** *p* <.001

Pairwise comparisons show that the differences in WWR for several conditions were significant. The difference in WWR between the 5C and 5R condition was the most significant. Participants obtained a significantly higher WWR score on the 5C format than they did on the 5R format. Other significant differences in WWR were also found between the 5C and 2C, 5C and 2R, and 5R and S conditions. Participants scored significantly higher on the 5C format than they did on either the 2C or 2R format. Participants scored significantly higher on the S format than they did on the 5R format.

Correct Letter Sounds per Minute

A repeated measures ANOVA showed that for the 53 participants in this sample the differences in CLS/min between the five test format conditions were statistically significant, *F*(4,49) = 11.677, *p* = .000 (p<.0005), partial η2 =.488. Based on guidelines proposed by Cohen (1988) this result suggests a large effect size. To determine the differences in CLS/min between each test format condition pairwise comparisons were completed to identify the significance between each test format condition. The results of the pairwise comparisons are presented in Table 7.

Table 7: Mean Differences in Pairwise Comparisons for CLS/min (Standard Error in Parentheses)

Test Format Condition	5R	2C	2R	S
5C	3.53 (2.50)	13.34 (2.68)***	12.93 (2.32)***	20.99 (3.28)***
5R		9.81 (2.22)***	9.40 (2.08)***	17.46 (2.74)***
2C			-0.42 (1.74)	7.65 (1.82)***
2R				8.06 (1.63)***

Note. 5C= 5 word column test format; 5R = 5 word row test format; 2C = 2 word column test format; 2R = 2 word row test format; S = single word test format.

*** *p* <.001

Most of the pairwise comparisons for CLS/min displayed significant differences. Participants obtained significantly higher CLS/min scores in the 5C compared to all other formats, except 5R. Participants obtained significantly higher CLS/min scores in the 5R format compared to the 2C, 2R, and S formats. Participant scored significantly higher in the 2C and 2R formats compared to the S format. Participant CLS/min scores were not significantly different between the 5C and 5R conditions, and the 2C and 2R conditions.

Whole Words Read per Minute

A repeated measures ANOVA showed that for the 53 participants in this sample the differences in WWR/min between the five test format conditions were statistically significant, *F*(4,49) = 7.913, *p* = .000 (*p*<.001), partial η2 =.392. Based on guidelines proposed by Cohen (1988) this result suggests a large effect size. To determine the differences in WWR/min between each test format condition pairwise comparisons were completed to identify the significance between each test format condition. The results of the pairwise comparisons are presented in Table 8.

Table 8: Mean Differences in Pairwise Comparisons for WWR/min (Standard Error in Parentheses)

Test Format Condition	5R	2C	2R	S
5C	2.63 (0.91)**	4.56 (1.00)***	4.31 (0.85)***	5.82 (1.08)***
5R		1.93 (0.84)*	1.69 (0.75)*	3.19 (0.89)**
2C			-0.24 (0.66)	1.26 (0.66)
2R				1.50 (0.75)*

Note. 5C= 5 word column test format; 5R = 5 word row test format; 2C = 2 word column test format; 2R = 2 word row test format; S = single word test format.

* *p* <.05; ** *p* <.01; *** *p* <.001

Most of the pairwise comparisons for WWR/min displayed significant differences. Participants scored significantly higher WWR/min in the 5C format compared to all other formats. Participants scored significantly higher WWR/min in the 5R format compared to the 2C, 2R, and S formats. Participant WWR/min scores were not significantly different between the 2C and 2R formats or 2C and S, and 2R and S format comparisons.

Research Question 2

Participants were asked to indicate their preference for the decoding words format. Nineteen participants selected the S format, fifteen selected the 5C format, nine selected the 5R format, six selected the 2C format, and four selected the 2R format. In the sample, 24.5% (n=13) of participants selected the format on which their CLS and CLS/min scores were highest. In the sample, 22.6% (n=12) of participants selected the format on which their WWR and WWR/min scores were highest.

A series of repeated measures ANOVAs were conducted to compare the effect of preference for test format on accuracy and fluency performance in the five test format conditions. The differences in CLS between the preferred format and non-preferred formats were examined using a repeated measures ANOVA. For this sample, results were not statistically significant, $F(4,49) = .374$, $p = .826$, partial η2 =.03. This result suggests that student preference for format had no effect on CLS performance.

The differences in WWR between the preferred format and non-preferred formats were examined using a repeated measures ANOVA. For this sample, results were not statistically significant, $F(4,49) = .834$, $p = .510$, partial η2 =.064. This result suggests that student preference for format had no effect on WWR performance.

The differences in CLS/min performance between preferred format and non-preferred formats were examined using a repeated measures ANOVA. For this sample, results were not statistically significant, $F(4,49) = .620$, $p = .650$, partial η2 =.048. This result suggests that student preference for format had no effect on CLS/min performance.

The differences in WWR/min performance between preferred format and non-preferred formats were examined using a repeated measures ANOVA. For this sample, results were not statistically significant, $F(4,49) = .314$, $p = .867$, partial η2 =.025. This result suggests that student preference for format had no effect on WWR/min performance.

Additional analyses of preference data sought to identify any relation between preference selection and performance on several criterion variables. Point biserial correlations were calculated between preference selection (coded 1: preferred and coded 0: not preferred) and WRC on DIBELS ORF and the GRADE total test score. None of the calculated correlations were significant. This result suggests that, in this sample, there is no relation between performance on those criterion measures and preference selection.

Research Question 3

To examine the evidence of criterion-related validity for the different decoding task formats study participants completed several decoding and reading tasks. Performance on each of the criterion measures was correlated with the four dependent variables (i.e., CLS, WWR, CLS/min, and WWR/min) for each of the five decoding test format conditions. Results are presented in tables 9-12.

Table 9: Correlations between Criterion Variables and CLS Test Format Performance (n=53)

Variable	5C CLS	5R CLS	2C CLS	2R CLS	S CLS
	r	*r*	*r*	*r*	*r*
NWF					
CLS	.504***	.495***	.549***	.574***	.644***
WWR	.560***	.534***	.572***	.633***	.674***
DIIP-S					
WRC	.592***	.490***	.617***	.625***	.714***
WRC/min	.454**	.384**	.458**	.499***	.584***
ORF					
WRC	.467***	.276*	.496***	.493***	.557***
GRADE					
Vocabulary Composite	.556***	.421**	.629***	.621***	.620***
Comprehension Composite	.382**	.176	.439**	.444**	.475***
Total Test Score	.497***	.306*	.566***	.566***	.585***

Note; NWF= DIBELS Next Nonsense Word Fluency; DIIP-S = Decoding Inventory for Instructional Planning – Screener; ORF= DIBELS Next Oral Reading Fluency.

* *p* <.05; ** *p* <.01; *** *p* <.001

Table 10: Correlations between Criterion Variables and WWR Test Format Performance (n=53)

Variable	5C WWR	5R WWR	2C WWR	2R WWR	S WWR
	r	*r*	*r*	*r*	*r*
NWF					
CLS	.540***	.583***	.550***	.541***	.612***
WWR	.683***	.728***	.695***	.695***	.744***
DIIP-S					
WRC	.715***	.678***	.713***	.710***	.770***
WRC/min	.541***	.529***	.515***	.535***	.600***
ORF					
WRC	.517***	.409**	.506***	.499***	.562***
GRADE					
Vocabulary Composite	.592***	.541***	.614***	.620***	.628***
Comprehension Composite	.363**	.227	.355**	.370**	.408**
Total Test Score	.502***	.393**	.507***	.520***	.547***

Note. NWF= DIBELS Next Nonsense Word Fluency; DIIP-S = Decoding Inventory for Instructional Planning – Screener; ORF= DIBELS Next Oral Reading Fluency.

* *p* <.05; ** *p* <.01; *** *p* <.001

Table 11: Correlations between Criterion Variables and CLS/min Test Format Performance (n=53)

Variable	5C	5R	2C	2R	S
	CLS/min	CLS/min	CLS/min	CLS/min	CLS/min
	r	*r*	*r*	*r*	*r*
NWF					
CLS	.787***	.824***	.737***	.826***	.784***
WWR	.789***	.825***	.752***	.840***	.833***
DIIP-S					
WRC	.715***	.698***	.751***	.780***	.802***
WRC/min	.800***	.783***	.752***	.821***	.752***
ORF					
WRC	.691***	.643***	.763***	.726***	.696***
GRADE					
Vocabulary Composite	.582***	.556***	.618***	.647***	.653***
Comprehension Composite	.655***	.563***	.739***	.677***	.596***
Total Test Score	.679***	.610***	.748***	.724***	.676***

Note. NWF= DIBELS Next Nonsense Word Fluency; DIIP-S = Decoding Inventory for Instructional Planning – Screener; ORF= DIBELS Next Oral Reading Fluency.

*** *p* <.001

Table 12: Correlations between Criterion Variables and WWR/min Test Format Performance (n=53)

Variable	5C WWR/min	5R WWR/min	2C WWR/min	2R WWR/min	S WWR/min
	r	*r*	*r*	*r*	*r*
NWF					
CLS	.753***	.790***	.695***	.761***	.759***
WWR	.804***	.855***	.780***	.846***	.854***
DIIP-S					
WRC	.765***	.747***	.788***	.820***	.816***
WRC/min	.780***	.762***	.712***	.779***	.775***
ORF					
WRC	.691***	.598***	.717***	.693***	.660***
GRADE					
Vocabulary Composite	.614***	.567***	.637***	.664***	.610***
Comprehension Composite	.613***	.476***	.634***	.598***	.539***
Total Test Score	.668***	.561***	.691***	.682***	.620***

Note. NWF= DIBELS Next Nonsense Word Fluency; DIIP-S = Decoding Inventory for Instructional Planning – Screener; ORF= DIBELS Next Oral Reading Fluency.

*** *p* <.001

In general, moderate to strong correlations were found between performance on the test format task scores and the criterion measures. The lowest correlation for each of the test format dependent variables (i.e., CLS, WWR, CLS/min and WWR/min) occurred in the 5R condition. The overall minimum correlation was *r* = .176 for the correlation between 5R CLS and the GRADE Comprehension Composite scores. The overall maximum correlation was *r* = .855, *p* < .001, between 5R WWR/min scores and NWF WWR scores. In general, stronger correlations were found between the test format fluency scores (i.e., CLS/min and WWR/min) and criterion variables than were found between the test format accuracy scores (i.e., CLS and WWR) and criterion variables.

Correlations between CLS performances on the test format task to criterion measures resulted correlations that ranged from trivial (*r* = .176 for the correlation between 5R CLS and the GRADE Comprehension Composite scores) to strong (*r* = .714, *p* <.001, for the correlation between S CLS and DIIP-S WRC scores).

Correlations between WWR performances on the test format task to criterion measures resulted in correlations that ranged from weak (*r* = .227 between 5R WWR and GRADE Comprehension Composite Scores) to strong (*r* = .770, *p* < .001, between S WWR and DIIP-S WRC scores).

Correlations between CLS/min performances on the test format task to criterion measures resulted in correlations that ranged from moderate (*r* = .556, *p* <.001, between 5R CLS/min scores and GRADE Vocabulary Composite scores) to very strong (*r* = .840, *p* < .001 between 2R CLS/min scores and NWF WWR scores).

Correlations between WWR/min performances on the test format task to criterion measures resulted in correlations that ranged from weak (*r* = .476, *p* <.001, between 5R

WWR/min and GRADE Comprehension Composite scores) to very strong (*r* = .855, *p* < .001, between 5R WWR/min scores and NWF WWR scores).

Correlation Comparisons Within Formats

To compare the correlation coefficients found in the first step of the correlation analyses, Meng's Z-test was run on all combinations of correlations. Results of Meng's Z-test for correlation comparisons amongst all the criterion variables are presented in Appendices A-D. Twenty-eight comparisons were completed for each of the test format conditions for each of the dependent variables for a total of 560 comparisons. Due to the large number of comparisons a Bonferroni correction was calculated. Forty-three of the comparisons were significant at the Bonferroni-corrected p-value of .001. The comparisons that were significant at *p* < .001 are those with differences that are unlikely to be attributed to chance occurrence.

No test format CLS comparisons were significant at *p* < .001. This indicates that the correlations between test format CLS scores (on any test format) and all criterion variables did not differ significantly. In addition, only two test format CLS/min comparisons were significant at *p* < .001. These comparisons were both found in the 5R CLS/min comparisons. The correlation between 5R CLS/min scores and NWF CLS scores was significantly stronger (Meng's Z = 3.24, df = 50, *p* = .001) than the correlation between 5R CLS/min scores and GRADE Vocabulary Composite scores. The other significant comparison showed that the correlation between 5R CLS/min scores and NWF WWR scores was significantly stronger (Meng's Z = 3.41, df = 50, *p* = .001) than the correlation between 5R CLS/min scores and GRADE Vocabulary Composite Scores.

This indicates that 5R CLS/min scores have a weaker relation to GRADE Vocabulary Composite scores than they do to NWF scores.

In contrast to the few significant comparisons found with CLS/min scores, 25 test format WWR comparisons were significant at p < .001. This indicates that there are differences in the strength of the relation between test format WWR scores and various criterion variable scores. Each significant comparison, and the test format where it occurred is included in Table 13. Each included comparison indicates which of the correlations is stronger. For example, the first comparison may be interpreted as the correlation between the 5C WWR scores and the NWF WWR scores is significantly stronger than the correlation between the 5C WWR scores and the NWF CLS scores.

Table 13: Significant WWR Score and Criterion Variable Comparisons (*p* < .001)

Comparison	5C	5R	2C	2R	S
NWF WWR > NWF CLS	✓	✓	✓	✓	✓
NWF WWR > DIIP-S WRC/min		✓			
NWF WWR > ORF WRC		✓			
NWF WWR > GRADE Comp.		✓			✓
NWF WWR > GRADE Total		✓			
DIIP-S WRC > DIIP-S WRC/min			✓		✓
DIIP-S WRC > ORF WRC		✓			✓
DIIP-S WRC > GRADE Comp.	✓	✓	✓	✓	✓
DIIP-S WRC > GRADE Total		✓			✓
GRADE Total > GRADE Comp.		✓	✓	✓	✓
Total Number of Significant Comparisons	2	9	4	3	7

Note. NWF= DIBELS Next Nonsense Word Fluency; DIIP-S = Decoding Inventory for Instructional Planning – Screener; ORF= DIBELS Next Oral Reading Fluency; CLS = Correct Letter Sounds; WWR = Whole Words Read; WRC = Words Read Correctly; Comp. = Comprehension Composite Score.

Last, 16 test format WWR/min comparisons were found to be significant at *p* < .001. This indicates that there are differences in the strength of the relation between test format WWR/min scores and various criterion variable scores. Each significant comparison, and the test format where it occurred is included in Table 14. Each included comparison indicates which of the correlations is stronger.

Table 14: Significant WWR/min Score and Criterion Variable Comparisons (*p* < .001)

Comparison	5C	5R	2C	2R	S
NWF CLS > GRADE Comp.		✓			
NWF WWR > NWF CLS				✓	✓
NWF WWR > ORF WRC		✓			
NWF WWR > GRADE Vocab.		✓			✓
NWF WWR > GRADE Comp.		✓			✓
NWF WWR > GRADE Total		✓			✓
DIIP-S WRC > GRADE Vocab.					✓
DIIP-S WRC > GRADE Comp.		✓		✓	✓
DIIP-S WRC > GRADE Total					✓
DIIP-S WRC/min > Comp.		✓			
Total Number of Significant Comparisons	0	7	0	2	7

Note. NWF= DIBELS Next Nonsense Word Fluency; DIIP-S = Decoding Inventory for Instructional Planning – Screener; ORF= DIBELS Next Oral Reading Fluency; CLS = Correct Letter Sounds; WWR = Whole Words Read; WRC = Words Read Correctly; Comp. = Comprehension Composite Score.

Correlation Comparisons Across Formats

The last step in examining evidence of criterion related validity of the test format task was to determine if any of the test formats displayed significant differences in their correlations with each criterion variable. To do this Meng’s Z test for comparing correlated correlation coefficients was done to compare the correlation of each test format dependent variable (i.e., CLS, WWR, CLS/min and WWR/min) with each criterion variable. Results of Meng's Z test across test formats are presented in Appendix E. Ten comparisons were completed per criterion variable. To determine significance a Bonferroni correction was calculated and resulted in an adjusted p-value of .005. Differences that are significant at the p < .005 are considered unlikely to be attributed to chance occurrence. Five comparisons were found to be significant using the Bonferroni-corrected *p* < .005. First, the correlation between S CLS scores and DIIP-S WRC scores was found to be stronger than the correlation between 5R CLS scores and DIIP-S WRC scores (Meng's Z = 2.83, df = 50, *p* = .005). Next, the correlation between S CLS scores and ORF WRC scores was found to be significantly stronger than the correlation between 5R CLS scores and ORF WRC scores (Meng's Z = 3.01, df= 50, *p* = .003). Then, the correlation between S CLS scores and GRADE Comprehension Composite scores was found to be significantly stronger than the correlation between 5R CLS scores and GRADE Comprehension Composite scores (Meng's Z = 3.05, df= 50, *p* = .002). Next, the correlation between S CLS scores and GRADE Total test scores was found to be significantly stronger than the correlation between 5R CLS scores and GRADE Total test scores (Meng's Z = 3.04, df= 50, *p* = .002). In all of the preceding significant differences the S CLS scores had a stronger correlation with the criterion variables than the 5R CLS scores did. The last significant difference found that the 2C CLS/min scores had a stronger correlation with the GRADE Comprehension Composite scores than the 2R CLS/min scores had with the GRADE Comprehension Composite scores (Meng's Z = 2.98, df= 50, *p* = .003). All other comparisons did not result in significant differences, indicating that for those comparisons there were no differences in how test format related to performance on the criterion variable measures.

CHAPTER 5

DISCUSSION

Overview

The purpose of this study was to examine the effects of test format on student decoding performance. The study was designed to answer the following research questions: (1) Does test format affect the decoding accuracy and fluency in second grade students? (2) Does student preference for test format affect the decoding accuracy and fluency in second grade students? (3) How do decoding tasks, presented in different formats, relate to other measures of reading? A repeated measures design was used to answer the research questions.

Rationale for Current Study

This study sought to expand the limited research on the effects of decoding test format on student performance. Decoding was selected as the content area for this study for a variety of reasons. First, most students who struggle in reading have decoding skill deficits (Moats & Tolman, 2009; Shaywitz, 2003; Torgesen, 2000). Next, decoding skill proficiency is strongly predictive of overall reading proficiency (Fletcher, Lyon, Fuchs & Barnes, 2007). Therefore, decoding is an important skill area to study in relation to test development. One way that teachers can determine if their students have learned decoding skills is through testing. Effective testing practices can have a positive effect on student outcomes, as teachers are able to use data acquired via testing to inform their instructional decision-making (Hosp & Ardoin, 2008). In addition, studies have shown that improving testing practices is associated with improving student outcomes (Black & Wiliam, 2009). It is therefore prudent that decoding tests be examined for possible improvements. The current study sought to explore if test format is a possible variable that may inform decoding test practices.

Research Question 1

Does test format affect the decoding accuracy and fluency in second grade students?

Previous research supports that list reading (i.e., reading individual words without context) is an adequate task for detecting reading impairments (Jenkins, Fuchs, Van den Broek, Espin, & Deno, 2003). This study examined how the format of such a list reading task may affect student performance. Results indicate that there are differences in accuracy and fluency performance between test formats. Participants completed a modified NWF task divided into five test format conditions (i.e., 5C, 5R, 2C, 2R, and S). Accuracy scores (i.e., CLS and WWR) and fluency scores (i.e., CLS/min and WWR/min) were calculated for each format. Comparisons of each score metric across all formats revealed performance differences depended on the score of interest as well as the test format. Significant differences in performance between test formats were found in all score metrics. However, results show that there were more differences in fluency score performance than in accuracy performance scores across test formats. In addition, the magnitude of the differences found was larger in fluency score comparisons than it was in accuracy score comparisons (i.e., CLS/min partial η2 =.488 and WWR/min partial η2 = .392 compared to CLS partial η2 = .177 and WWR partial η2 = .219). These data suggest that test format may have a larger effect on fluency than on accuracy.

The numbers of words presented at a time during the task may impact fluency performance. This was demonstrated as fluency scores on the 5C and 5R formats were significantly higher compared to the fluency scores on the 2C, 2R, and S formats. Additional evidence that the number of words impacts fluency performance is that participant CLS and WWR scores were significantly higher on the S format than on the 5R format. However, when those accuracy scores were converted to fluency scores (i.e., CLS/min and WWR/min) then participants scored significantly higher on the 5R format compared to the S format. These data align with results of previous studies of preview benefit in reading and eye tracking showing that readers who are exposed to multiple words at a time are able to read them faster than when presented with words in isolation (Chace, Rayner, & Well, 2005). Readers are able to read faster when they are presented with multiple words at a time, as while they are fixated on one word they are also able to begin processing the next word. In contrast, when only presented with a single word at a time there is no opportunity for preview benefit as each word is presented as a totally isolated reading task.

Results also show that there are no significant performance differences on any of the scoring metrics between 2C and 2R performances. In contrast, participant performance on the 5C format (in all score metrics except CLS/min) was significantly higher than performance on the 5R format. These data suggest that layout (i.e., columns and rows) may impact decoding performance. This result has implications for test development and may be related to a mode effect. The 5R test format is the format that is the closest to traditional paper-based reading tasks, but decoding scores observed in this study were generally, significantly higher in the 5C test format condition. In addition, prior to the test conditions reading task, all participants completed the DIBELS Next NWF benchmark, which has readers read nonsense words from left to right. In a sense, that task may have primed participants for the 5R test format, but results show that compared to other formats scores 5R scores were depressed. It is possible that the optimal format for paper and computer based tests are different. Further investigation into the role of a mode effect on decoding performance is necessary to determine optimal test format selections.

Research Question 2

Does student preference for test format affect the decoding accuracy and fluency in second grade students?

Results of this study do not indicate that student preference for format affected decoding performance. After completing the 100-word nonsense word reading task participants were asked to indicate which of the five formats they liked best. Comparing their performance on their selected preferred format to their performance on other formats yielded no significant differences on any of the score metrics. There were no significant accuracy or fluency performance differences between the participant selected preferred format compared to performance on non-preferred formats (i.e., partial η2 range = .025- .064). These data suggest that preference for format did not affect decoding performance on this task. While a percentage of participants preferred format selection did match their best performance the differences between their performance on their selected format and others was not significantly different.

There are several reasons why this result may have been observed. First, it is possible that due to the skill level of participants in this sample, performances on all formats was relatively equal (i.e., participant performance was consistent across all formats). Limited variation in performance between preferred and non-preferred formats would have impacted results. In addition, it is possible that participants in this sample, due to their age and proficiency with decoding skills were not self-aware of how test format affected their decoding performance, and therefore were less sensitive to how format affected their performance. While some participants remarked that they preferred reading the formats with multiple words because they could read them faster, others remarked that they preferred to read only one word at a time. It is possible that those participants that had mastered the decoding skills necessary were able to notice the observed fluency increases in the multi-word formats compared to those participants that were not as proficient and therefore did not notice performance differences between formats. However, because no relation was found between overall reading proficiency and format preference future studies will be necessary to determine if there is a relation between reading skill level and format preference.

In addition, while results of this study do not align with previous research on motivation and reading performance it is possible that the improved reading performance is more closely tied to control in reading choices than to physical formats of reading tasks. Previous studies have supported that student choice in selection of reading materials is associated with improved student engagement and performance on the selected material (Randi & Corno, 2000; Ciampa, 2016). This preference data collected in this study was not related to student choice in the actual task (i.e., participants were required to complete the task in all formats, and not just their preferred format). The preference data analyzed in this study asked participants to select a preferred test format after they had engaged in a decoding task that included five format options. It is possible that performance differences between formats would be detectable had participants been able to control their format choices at the onset of the task (i.e., increased performance on the format the participant selects to engage in rather than engaging with all possible formats). Additional research is necessary to determine how student preference for format may assist in optimizing test time.

Research Question 3

How do decoding tasks, presented in different formats, relate to other measures of reading?

Participants in this study completed a variety of decoding and reading tasks, some of which were very similar to the test format conditions task (i.e., DIBELS Next NWF) and some that were measures of broader reading (i.e., GRADE). By comparing the accuracy and fluency performance on the test format conditions task to performance on other measures of decoding and reading it was determined what evidence of criterion-related validity existed for each test format. Results of this study indicate that some evidence of criterion-related validity exists for each test format for both accuracy and fluency scoring metrics. However, the strength of the evidence of criterion-related validity depended on the score metric, criterion of interest, and test format. In addition, examination of the differences in correlations within each test format was used to determine if the evidence of criterion-related validity was dependent on the criterion variable of interest, and exanimation of the differences in correlations between test formats was used to determine if the evidence of criterion-related validity was dependent on test format.

Accuracy Scores

The correlations between CLS scores and criterion variables ranged from trivial to strong (*r* range = .176-.714). This indicates that some evidence of criterion-related validity exists for each test format based on CLS scores. Examining correlation comparisons within each test format resulted in no significant comparisons, indicating that CLS scores within any one format related to all criterion variables with about the same strength. For example, the 2C CLS score relates to DIBELS NWF WWR scores with about the same strength as they do to GRADE Comprehension composite scores. The same interpretation holds for all CLS scores regardless of test format, and for all criterion measures. These data provide evidence of criterion-related validity for CLS scores within each test format.

While no significant differences were found between the correlation of CLS performance and criterion performance within each test format, significant differences were found when comparing the correlations between CLS performance and criterion performance between formats. For CLS scores, significant differences were found between the S format and the 5R format for certain criterion variables. Results show that the correlation between S CLS scores and some criterion measures (i.e., DIIP-S WRC, ORF WRC, GRADE Comprehension composite, and GRADE total test scores) were significantly stronger than the correlation between the 5R CLS scores and those criterion variables. These data indicate that for these criterion variables the CLS performance on the S format is a better indicator than CLS performance on the 5R format. This result suggests that for CLS scores, a single word test format may be preferable compared to a 5R format in terms of criterion-related validity for certain criterion measures.

A different pattern of results was found in the analysis of WWR accuracy scores. The correlations between WWR scores and criterion performance scores ranged from weak to strong (*r* range = .227 - .770). This indicates some evidence of criterion-related validity exists for WWR scores on each test format. Within each test format, correlation comparisons resulted in 25 significant comparisons. This indicates that WWR scores were more strongly related to some criterion variables (e.g., DIBELS Next WWR scores and DIIP-S WRC scores) than to others (e.g., DIBELS Next CLS scores, GRADE Comprehension composite scores) within a test format. Therefore, the validity of the WWR score on any given format may be dependent on the criterion measure of interest. These results also demonstrate that test format does influence the relation between WWR score and criterion performance, as while significant differences were found on all test formats, some formats, notably 5R and S formats, had more significant correlation differences between WWR scores and criterion scores than other formats. The relation between WWR score and criterion measures of decoding and reading may differ depending on the test format. Therefore, assumptions of equivalent validity for various criterion measures across test formats are not supported by the results of this study. Results of this study support that evidence of criterion-related validity must be independently measured for every test format, and some formats may have stronger validity evidence than others.

While there were many significant differences in the relation between WWR scores and criterion scores within test formats, there were no significant differences between test formats. This result shows there is no one test format WWR score that results in a significantly stronger relation to any of the criterion scores compared to other test formats. Therefore, using WWR on a variety of test formats may result in differences in how well the score relates to different criterion measures of decoding and reading, but there is no format that relates to these criterion measures better than others.

Overall, results show that there is evidence of criterion-related validity for accuracy scores for all test formats. The strength of the evidence is dependent on the score metric of choice (i.e., CLS or WWR), the criterion measure of interest, as well as test format. WWR scores on different test formats resulted in more variation in correlation comparisons within test formats than did CLS scores. It is important to note that WWR scores requires application of a more advanced decoding skill (i.e., the ability to apply letter sounds to multiple letters and accurately blend all phonemes together). Because participants in this sample were in the beginning of second grade there is likely a distribution of decoding abilities, with some students having mastered basic decoding skills and others that have yet to master the decoding skills necessary to read whole nonsense words. This may have resulted in inconsistent whole word reading performance across the nonsense word reading task and increased variation in the WWR scoring metric. Thus within test formats results show that WWR scores relate differently to various criterion measures of decoding and reading, while CLS scores, being a more basic decoding skill, were likely more consistent across all performance opportunities (i.e., test formats), and therefore resulted in no difference relation to criterion performance scores. In terms of test optimization and efficiency, results show that all formats have some evidence of criterion-related validity, but the S format outperforms the 5R format for CLS score metric.

Fluency Scores

In general, the correlations between the fluency score metrics and the criterion measures were higher than the correlations found in the accuracy score metric analyses. The correlations between CLS/min scores and criterion measure performance scores ranged from moderate to very strong (*r* range = .556-.840). This indicates that there is some evidence of criterion-related validity for CLS/min scores for all test formats. In contrast to no significant comparisons within the CLS score metric, CLS/min correlation comparisons within each test format and all criterion measures resulted in two significant comparisons. In the 5R format, the correlation between CLS/min score and NWF CLS score is significantly stronger than the correlation between CLS/min score and GRADE Vocabulary score. Also in the 5R format, the correlation between CLS/min score and NWF WWR score is significantly stronger than the correlation between CLS/min score and GRADE Vocabulary composite score. No other correlation comparisons were significant, which indicates that within the other test formats CLS/min scores relate to all criterion scores with about the same strength.

Comparing the correlations between test formats resulted in only one significant comparison. This is a different pattern than what was found in the CLS score metric comparisons across formats, where there were four significant comparisons that all involved the S and 5R formats. In the CLS/min score comparisons, the 2C CLS/min score is more strongly correlated with the GRADE Comprehension composite score than the 2R CLS/min score. This result indicates that if the GRADE Comprehension composite score is the criterion of interest, and CLS/min is the score metric being used, then the 2C test format may be a more preferred test format in terms of criterion-related validity than the 2R test format. No other significant comparisons were found. This indicates that all the other test format CLS/min scores related to the criterion measure scores with about the same strength.

Following the pattern observed in accuracy score analyses, where more significant differences were found in the WWR score metric than in the CLS score metric, there was also more variation in the results of WWR/min fluency score analyses than there were in CLS/min score analyses. Correlations between WWR/min scores and criterion scores ranged from weak to very strong (*r* range = .476-.855). This provides some evidence of criterion-related validity for WWR/min scores for all test formats. Correlation comparisons within each test format resulted in 16 significant differences indicating that within a format the WWR/min scores related differently to various criterion measure scores. However, not all test formats had significant differences. WWR/min scores on the 5C and 2C test formats relate to all criterion measures the same. By contrast, within 5R, 2R, and S formats there are significant differences in how WWR/min scores relate to various criterion scores. This result is important as within a given format strength of the validity evidence depends on the criterion of interest. In addition, this result is an important indicator that test format does influence evidence of validity, as not all test formats included significant differences. Therefore, test format must be a variable studied in test development and validity studies. This study sought to determine if test format affected student performance and if test format may be a variable that could improve decoding test practices. Of critical consideration in test development is the psychometric qualities of the test. Providing evidence of criterion-related validity for any task is a necessary component of test development, and results from this study show that the strength of the evidence of criterion-related validity depends on the test format of the task. This study provides evidence that assuming equivalent validity across test formats is not advised.

Comparing correlations of WWR/min scores to criterion variables across test formats yielded no significant differences. Between formats there was no difference in how WWR/min scores related to any of the criterion measure scores. This result indicates that WWR/min scores on all test formats relate to any of the criterion measures the same. Therefore, in terms of criterion-related validity there is no optimal format for the WWR/min score metric.

Overall, results show that there is evidence of criterion-related validity for fluency scores for all test formats. In general, the correlations between fluency scores and criterion scores were stronger than those found in the accuracy score analyses, indicating that fluency may be a better score metric for predicting criterion performance than the included accuracy scores. Measuring fluency of decoding skills is important, as laborious or inefficient application of decoding accuracy skills is not sufficient for overall reading improvement (Joshi & Aaron, 2002). However, as with the accuracy score metrics, the strength of the evidence is dependent on the score metric of choice (i.e., CLS/min or WWR/min), the criterion measure of interest, as well as test format. As such, all three aspects should be considered during test development.

Implications

The purpose of this study was to examine the effects of test format on decoding accuracy and fluency. Time that teachers allocate for testing limits the classroom time that is available for instruction. Therefore, it is prudent to examine possible ways to maximize testing processes. This study examined possible variables associated with the design of the testing task that might improve the efficiency of the task. Results of this study have several implications for test development. Overall, results support that test format is a variable that should be examined during test development, and multiple test formats may be valid options for test design depending on the purpose of the test. The performance metric of interest, whether it is accuracy or fluency, may inform format choices in terms of the number of words that are presented to a student at a time. Test format demonstrated a greater effect on fluency scores than it did on accuracy scores. This result implies that the optimal format of the test, even for the same task, may change depending on the score of interest. For example, when a measure is focused on accuracy the optimal format for the test may be a single word presentation, while a multiple word format may be more appropriate when fluency becomes the target of the measure. When introducing new skills, teachers are likely to focus on developing accuracy of that skill. After sufficient practice of the skill, instruction will transition to fluency or application of that skill in context. Teachers in the primary grades of elementary school (e.g., K-1) are likely providing focused instruction on building decoding skill accuracy. Second grade serves as a transition point for decoding skills as the focus of instruction transitions to a fluency focus (Chall, 1996). During the introduction of new skills, accuracy is likely the performance metric of interest for teachers, while fluency becomes of interest during the application and generalization phase of instruction. In addition, the evidence of criterion-related validity that was found in this study supports that multiple test formats for decoding can be valid options for test design. Results of this study support that test formats, which are traditionally static layouts (i.e., single published layouts) can be dynamic and altered for the purpose of the test.

In addition to the score metric of interest informing optimal test format, results of this study have particular implications for the format of screening measures. The results of the criterion-related validity analyses show that within a test format there are differences in how well results relate to various criterion measures. Therefore, it is critical that when designing screening tests that are predictive of an outcome criterion that the criterion of interest inform the optimal test format. Some test formats and score metrics are more strongly related to some criterion variables than others, and these correlations should be examined and considered in validity studies for decoding tests. For example, if the outcome measure of interest is comprehension (e.g., GRADE Comprehension composite scores) then, based on the results of this study, a 2C format is a more optimal format selection than a 2R format selection. Results show that the strength of the relation to the criterion of interest depends on both the test format as well as the scoring metric used. When selecting a test format for a screening measure it is important to consider the criterion of interest.

Overall, results of this study show that there are multiple viable options for test format in a decoding task. The selection of a test format should consider the performance metric of interest (i.e., accuracy or fluency scores) as well as evidence of criterion-related validity. Test format does affect student decoding performance and should be studied during test construction and development. Results of this study support that test format may be a variable that can maximize test efficiency in classrooms.

Limitations

There are several limitations to this study. First, there are limitations in the representativeness of the sample of participants in this study. The sample had a lack of diversity of ethnicities, disability statuses, and language learners. This makes it difficult to generalize results to broader populations. In addition, students in this sample were not participating in decoding or phonics instruction in their classrooms, and most had limited previous exposure to explicit phonics instruction. This certainly may have affected performance on the decoding tasks completed during the study. Therefore, differences in performance on test formats may have been influenced by novelty of the actual task, and not by the test format. Replication of results and future studies of the effects of test format are needed to confirm and define the effect of test format on decoding performance.

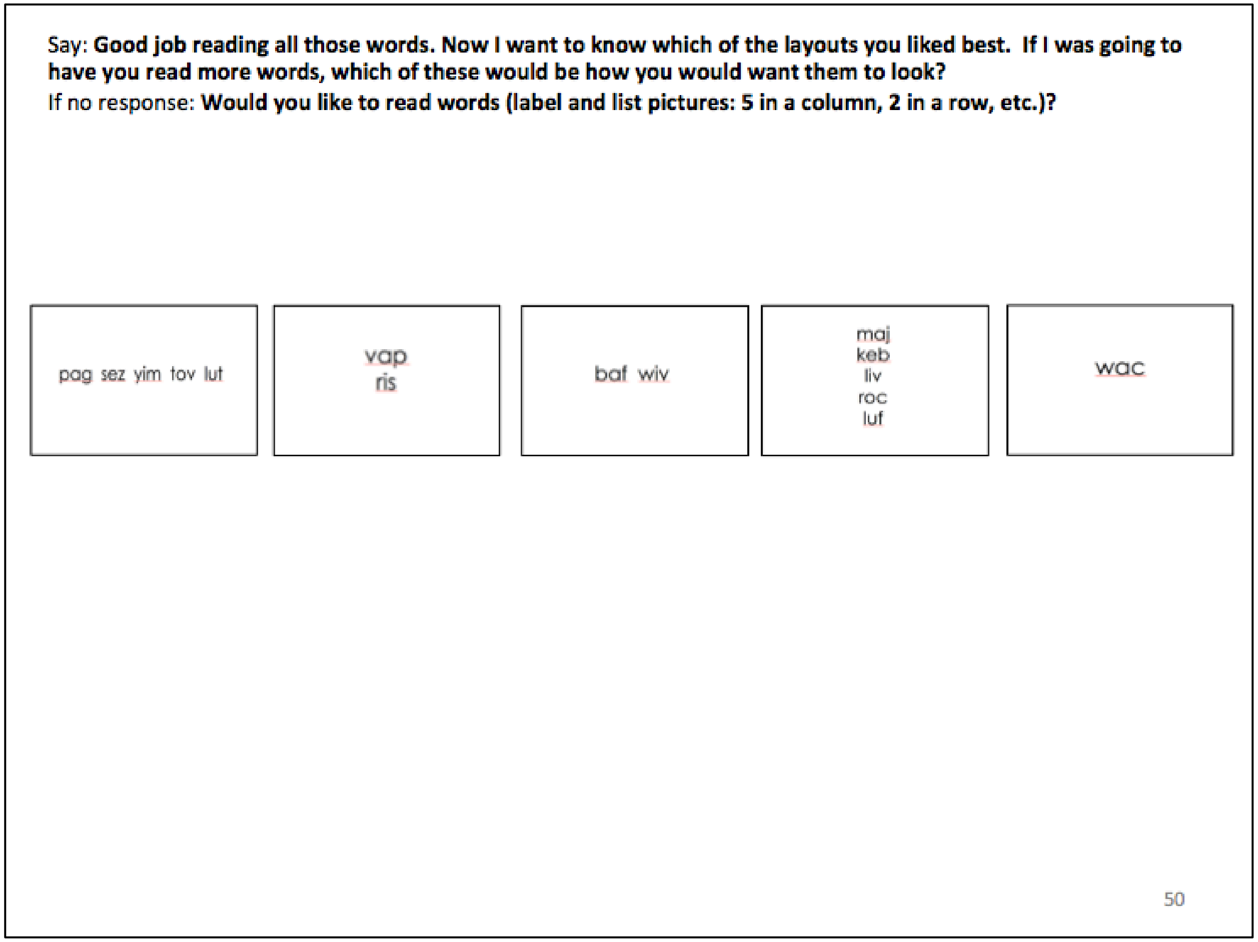
Future Directions

Results of this study show that test format does affect student performance on a computer based decoding task. Future studies are needed to determine optimal test format designs for decoding tasks beyond CVC word patterns. In addition, while this study focused on test format for decoding skills there is a need to examine ways to maximize test efficiency across other decoding and broader literacy skills. With the likely continued growth in computer based testing there is also a need to examine any mode effect that results from reading on a technology platform. The differences in column and row reading performance in this study may be indicative that students change their reading habits based on mode (i.e., they read differently on paper (i.e., left to right) than they do on a screen (i.e., top to bottom). In addition, the influence of student choice and control as facets of motivation in testing situations requires further study. Examination of the role of control in test presentation may be an additional option for maximizing test efficiency in classrooms.

In conclusion, this study served to collect preliminary data on the effects of test format on the decoding performance of second grade students. Results support that test format does affect accuracy and fluency performance. In addition, examination of the relation between decoding performance on the test format task and other measures of decoding and reading indicate that test format also affects the strength of the evidence of validity of the decoding task. These results support the hypothesis that test format is a potential variable in exploring ways to improve the efficiency of decoding testing practices. While additional research is necessary to explore the nuances of test optimization, with particular attention paid to the role of computer based tests, this study does expand the limited research on the effects of test format and highlights the potential for improved test designs that improve test efficiency in practice.

APPENDIX A

EXAMPLE OF PREFERENCE SELECTION SLIDE



Say: **Good job reading all those words. Now I want to know which of the layouts you liked best. If I was going to have you read more words, which of these would be how you would want them to look?**

If no response: **Would you like to read words (label and list pictures: 5 in a column, 2 in a row, etc.)?**

pag sez yim tov lut

vap
ris

baf wiv

maj
keb
liv
roc
luf

wac

50

APPENDIX B

Z SCORES AND SIGNIFICANCE VALUES FROM MENG'S Z TEST FOR CLS FOR CRITERION VARIABLES

Criterion Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent Variable	WWR Z (*p)*	WRC Z (*p)*	WRC/min Z (*p)*	WRC Z (*p)*	Vocab. Z (*p)*	Comp. Z (*p)*	Total Score Z (*p)*
				5C				
NWF	CLS	1.55 (.121)	1.03 (.305)	.73 (.464)	.36 (.718)	.48 (.633)	1.01 (.314)	.11 (.913)
	WWR		.45 (.655)	1.55 (.121)	.95 (.343)	.04 (.968)	1.49 (.135)	.61 (.544)
DIIP-S	WRC			2.18 (.030)	1.68 (.093)	.48 (.629)	2.23 (.026)	1.28 (.200)
ORF	WRC/min				.17 (.862)	1.00 (.317)	.75 (.452)	.49 (.628)
GRADE	WRC					1.04 (.296)	1.50 (.134)	.57 (.572)
	Vocab.						1.79 (.074)	1.06 (.290)
	Comp.							2.62 (.009)

(continued)

Criterion Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
	Variable	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
				5R				
NWF	CLS	1.06 (.287)	.06 (.956)	1.59 (.113)	2.00 (.045)	.64 (.524)	2.48 (.013)	1.62 (.106)
	WWR		.58 (.562)	2.12 (.034)	2.44 (.015)	1.05 (.295)	2.79 (.005)	2.02 (.043)
DIIP-S	WRC			1.57 (.116)	2.61 (.009)	.84 (.401)	3.04 (.002)	2.24 (.025)
	WRC/min				1.35 (.178)	.34 (.736)	2.04 (.042)	.82 (.414)
ORF	WRC					1.54 (.124)	1.63 (.102)	.52 (.606)
GRADE	Vocab.						2.29 (.022)	1.87 (.062)
	Comp.							2.74 (.006)

(continued)

Criterion Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent Variable	WWR Z (*p)*	WRC Z (*p)*	WRC/min Z (*p)*	WRC Z (*p)*	Vocab. Z (*p)*	Comp. Z (*p)*	Total Score Z (*p)*
				2C				
NWF	CLS	.65 (.515)	.82 (.411)	1.36 (.175)	.53 (.594)	.79 (.433)	.95 (.344)	.17 (.869)
	WWR		.64 (.522)	1.68 (.093)	.79 (.430)	.60 (.550)	1.15 (.250)	.06 (.952)
DIIP-S	WRC			2.54 (.011)	1.67 (.096)	.17 (.865)	1.96 (.050)	.72 (.472)
	WRC/min				.51 (.607)	1.75 (.081)	.20 (.840)	1.26 (.209)
ORF	WRC					1.64 (.101)	1.03 (.302)	1.37 (.171)
GRADE	Vocab.						2.07 (.039)	1.20 (.229)
	Comp.							3.02 (.003)

(continued)

Criterion Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent Variable	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
		Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
2R								
NWF	CLS	1.74 (.082)	.63 (.530)	1.15 (.251)	.82 (.410)	.47 (.642)	1.14 (.256)	.08 (.937)
	WWR		.12 (.906)	2.07 (.039)	1.50 (.133)	.13 (.896)	1.70 (.090)	.70 (.485)
DIIP-S	WRC			2.05 (.040)	1.82 (.068)	.06 (.955)	2.00 (.045)	.84 (.403)
	WRC/min				.08 (.935)	1.26 (.206)	.60 (.552)	.79 (.429)
ORF	WRC					1.57 (.117)	.89 (.375)	1.42 (.154)
GRADE	Vocab.						1.92 (.055)	1.05 (.295)
	Comp.							2.90 (.004)

(continued)

Criterion								
Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
	Variable	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*
				S				
NWF	CLS	.94 (.348)	.96 (.336)	.99 (.322)	.95 (.341)	.25 (.803)	1.56 (.118)	.62 (.539)
	WWR		.65 (.516)	1.48 (.138)	1.34 (.182)	.61 (.545)	1.87 (.062)	.97 (.333)
DIIP-S	WRC			2.34 (.020)	2.37 (.018)	1.42 (.157)	2.84 (.005)	1.96 (.050)
	WRC/min				.39 (.694)	.39 (.696)	1.24 (.216)	.01 (.990)
ORF	WRC					.80 (.426)	1.54 (.125)	.57 (.572)
GRADE	Vocab.						1.59 (.111)	.67 (.502)
	Comp.							2.67 (.008)

Note. NWF = Nonsense Word Fluency; DIIP-S = Decoding Inventory for Instruction Planning – Screener; ORF= Oral Reading Fluency; CLS = Correct Letter Sounds; WWR = Whole Words Read; WRC = Words Read Correctly; Vocab. = Vocabulary; Comp. = Comprehension.

APPENDIX C

Z SCORES AND SIGNIFICANCE VALUES FROM MENG'S Z TEST FOR WWR FOR CRITERION VARIABLES

Criterion Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent Variable	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
		Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
				5C				
NWF	CLS	**4.27 (.000)**	2.25 (.024)	.02 (.988)	.23 (.816)	.50 (.620)	1.47 (.140)	.36 (.722)
	WWR		.52 (.600)	2.30 (.022)	1.86 (.062)	1.01 (.314)	2.86 (.004)	1.89 (.059)
DIIP-S	WRC			3.06 (.002)	2.93 (.003)	1.82 (.068)	**3.98 (.000)**	3.10 (.002)
	WRC/min				.34 (.736)	.53 (.596)	1.90 (.057)	.46 (.648)
ORF	WRC					.92 (.360)	2.75 (.006)	.29 (.773)
GRADE	Vocab.						2.37 (.018)	1.64 (.101)
	Comp.							3.16 (.002)

(continued)

Criterion Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent Variable	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
		Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
				5R				
NWF	CLS	**4.55 (.000)**	1.22 (.223)	.84 (.401)	1.72 (.086)	.40 (.690)	2.89 (.004)	1.73 (.083)
	WWR		.83 (.409)	**3.31 (.001)**	**3.52 (.000)**	2.08 (.037)	**4.38 (.000)**	**3.43 (.001)**
DIIP-S	WRC			2.54 (.011)	**3.70 (.000)**	1.92 (.055)	**4.75 (.000)**	**3.86 (.000)**
	WRC/min				1.62 (.106)	.12 (.904)	3.09 (.002)	1.52 (.128)
ORF	WRC					1.51 (.132)	3.06 (.002)	.29 (.774)
GRADE	Vocab.						3.06 (.002)	2.55 (.011)
	Comp.							**3.57 (.000)**

(continued)

Criterion Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
	Variable	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
				2C				
NWF	CLS	**4.38 (.000)**	2.11 (.035)	.53 (.594)	.45 (.656)	.62 (.534)	1.63 (.104)	.41 (.685)
	WWR		.30 (.766)	2.90 (.004)	2.13 (.033)	.92 (.358)	3.06 (.002)	1.98 (.047)
DIIP-S	WRC			**3.43 (.001)**	3.04 (.002)	1.49 (.137)	**4.03 (.000)**	3.00 (.003)
	WRC/min				.13 (.901)	1.03 (.303)	1.69 (.091)	.09 (.926)
ORF	WRC					1.33 (.185)	2.68 (.007)	.02 (.985)
GRADE	Vocab.						2.71 (.007)	1.98 (.048)
	Comp.							**3.45 (.001)**

(continued)

Criterion Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
	Variable	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
				2R				
NWF	CLS	**4.63 (.000)**	2.17 (.030)	.09 (.927)	.42 (.673)	.77 (.443)	1.43 (.153)	.20 (.843)
	WWR		.25 (.805)	2.60 (.009)	2.20 (.028)	.86 (.392)	2.94 (.003)	1.86 (.063)
DIIP-S	WRC			3.05 (.002)	3.08 (.002)	1.35 (.176)	**3.85 (.000)**	2.78 (.006)
	WRC/min				.50 (.616)	.90 (.370)	1.76 (.078)	.18 (860)
ORF	WRC					1.49 (137)	2.29 (.022)	.40 (.687)
GRADE	Vocab.						2.63 (.009)	1.86 (.062)
	Comp.							**3.44 (.001)**

(continued)

Criterion								
Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
	Variable	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
				S				
NWF	CLS	**4.26 (.000)**	2.25 (.025)	.20 (.845)	.54 (.591)	.16 (.870)	1.79 (.073)	.65 (.517)
	WWR		.48 (.634)	2.53 (.012)	2.21 (.027)	1.40 (.162)	**3.23 (.001)**	2.23 (.026)
DIIP-S	WRC			**3.24 (.001)**	**3.31 (.001)**	2.27 (.023)	**4.38 (.000)**	**3.49 (.001)**
	WRC/min				.56 (.575)	.31 (.758)	2.14 (.033)	.65 (.515)
ORF	WRC					.84 (.401)	2.83 (.005)	.30 (.765)
GRADE	Vocab.						2.36 (.018)	1.53 (.125)
	Comp.							**3.25 (.001)**

Note. NWF = Nonsense Word Fluency; DIIP-S = Decoding Inventory for Instruction Planning – Screener; ORF= Oral Reading Fluency; CLS = Correct Letter Sounds; WWR = Whole Words Read; WRC = Words Read Correctly; Vocab. = Vocabulary; Comp. = Comprehension.

Bold: *p*≤.001

APPENDIX D

Z SCORES AND SIGNIFICANCE VALUES FROM MENG’S Z TEST FOR CLS/MIN FOR CRITERION VARIABLES

Criterion Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
	Variable	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
				5C				
NWF	CLS	.08 (.939)	1.14 (.252)	.29 (.771)	1.35 (.178)	2.40 (.017)	1.62 (.106)	1.42 (.156)
	WWR		1.35 (.178)	.24 (.808)	1.40 (.161)	2.54 (.011)	1.64 (.102)	1.47 (.141)
DIIP-S	WRC			1.82 (.068)	.40 (.692)	1.96 (.050)	.80 (424)	.58 (.559)
	WRC/min				2.06 (.040)	2.80 (.005)	2.19 (.028)	1.93 (.053)
ORF	WRC					1.46 (.144)	.79 (.429)	.28 (.783)
GRADE	Vocab.						.87 (.386)	1.92 (.054)
	Comp.							.67 (.506)

(continued)

Criterion Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent Variable	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
		Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*
				5R				
NWF	CLS	.04 (.966)	2.08 (.038)	.95 (.344)	2.56 (.010)	**3.24 (.001)**	3.10 (.002)	2.78 (.006)
	WWR		2.39 (.017)	.95 (.340)	2.62 (.009)	**3.41 (.001)**	3.10 (.002)	2.84 (.005)
DIIP-S	WRC			1.76 (.078)	.86 (.388)	2.03 (.042)	1.66 (.096)	1.34 (.180)
	WRC/min				2.49 (.013)	2.79 (.005)	3.03 (.002)	2.55 (.011)
ORF	WRC					1.11 (.266)	1.62 (.105)	.71 (.481)
GRADE	Vocab.						.08 (.938)	1.02 (.309)
	Comp.							1.19 (.232)

(continued)

Criterion								
Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
	Variable	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
				2C				
NWF	CLS	.53 (.595)	.22 (.828)	.30 (.764)	.37 (.710)	1.35 (.178)	.03 (.980)	.15 (.884)
	WWR		.02 (.986)	*	.16 (.871)	1.62 (.106)	.17 (.867)	.06 (.956)
DIIP-S	WRC			.02 (.983)	.22 (.826)	2.07 (.038)	.18 (.857)	.05 (.957)
	WRC/min				.21 (.833)	1.67 (.096)	.20 (.841)	.07 (.949)
ORF	WRC					2.13 (.033)	.60 (.550)	.39 (.698)
GRADE	Vocab.						1.58 (.115)	2.77 (.006)
	Comp.							.28 (.780)

(continued)

Criterion Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent Variable	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
		Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
				2R				
NWF	CLS	.61 (.544)	.85 (.397)	.12 (.903)	1.56 (.118)	2.37 (.018)	2.01 (.045)	1.52 (.128)
	WWR		1.28 (.202)	.47 (.640)	1.86 (.063)	2.74 (.006)	2.26 (.024)	1.81 (.071)
DIIP-S	WRC			.97 (.332)	.99 (.325)	2.18 (.029)	1.50 (.133)	1.01 (.312)
	WRC/min				1.91 (.056)	2.44 (.015)	2.29 (.022)	1.68 (.094)
ORF	WRC					1.14 (.254)	1.12 (.262)	.05 (.961)
GRADE	Vocab.						.38 (.705)	1.64 (.101)
	Comp.							1.37 (.172)

(continued)

Criterion Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
	Variable	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*
				S				
NWF	CLS	1.99 (.046)	.32 (.749)	.67 (.500)	1.24 (.217)	1.66 (.097)	2.16 (.031)	1.41 (.160)
	WWR		.67 (.501)	1.79 (.073)	2.12 (.034)	2.53 (.011)	2.94 (.003)	2.26 (.024)
DIIP-S	WRC			1.12 (.264)	1.93 (.053)	2.52 (.012)	2.89 (.004)	2.24 (.025)
	WRC/min				1.01 (.315)	1.27 (.205)	2.13 (.033)	1.14 (.254)
ORF	WRC					.61 (.544)	2.13 (.033)	.46 (.645)
GRADE	Vocab.						.68 (.496)	.48 (.635)
	Comp.							2.14 (.032)

Note. * Correlations were equal and therefore the result of the Meng's Z-test was null. NWF = Nonsense Word Fluency; DIIP-S = Decoding Inventory for Instruction Planning – Screener; ORF= Oral Reading Fluency; CLS = Correct Letter Sounds; WWR = Whole Words Read; WRC = Words Read Correctly; Vocab. = Vocabulary; Comp. = Comprehension.

Bold: *p*≤.001

APPENDIX E

Z SCORES AND SIGNIFICANCE VALUES FROM MENG’S Z TEST FOR WWR/MIN FOR CRITERION VARIABLES

Criterion Measure		NWF	DIIP-S		ORF		GRADE	
	Dependent Variable	WWR Z (*p*)	WRC Z (*p*)	WRC/min Z (*p*)	WRC Z (*p*)	Vocab. Z (*p*)	Comp. Z (*p*)	Total Score Z (*p*)
				5C				
NWF	CLS	1.93 (.053)	.19 (.847)	.57 (.571)	.83 (.406)	1.60 (.110)	1.57 (.117)	1.05 (.293)
	WWR		.77 (.444)	.53 (.599)	1.65 (.098)	2.45 (.014)	2.29 (.022)	1.84 (.066)
DIIP-S	WRC			.33 (.741)	1.28 (.199)	2.38 (.017)	2.06 (.039)	1.64 (.102)
ORF	WRC/min				1.64 (.101)	2.13 (.033)	2.38 (.017)	1.62 (.105)
GRADE	WRC					1.05 (.292)	1.67 (.095)	.53 (.600)
	Vocab.						.01 (.991)	1.08 (.279)
	Comp.							1.48 (.139)

(continued)

Criterion Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
	Variable	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*
				5R				
NWF	CLS	2.75 (.006)	.71 (.476)	.60 (.548)	2.48 (.013)	2.59 (.010)	**3.32 (.001)**	2.71 (.007)
	WWR		2.26 (.024)	2.17 (.030)	**3.75 (.000)**	**3.89 (.000)**	**4.42 (.000)**	**3.94 (.000)**
DIIP-S	WRC			.32 (.751)	2.37 (.018)	2.70 (.007)	**3.32 (.001)**	2.87 (.004)
	WRC/min				2.77 (.006)	2.36 (.018)	**3.67 (.000)**	2.81 (.005)
ORF	WRC					.39 (.698)	2.33 (.020)	.75 (.451)
GRADE	Vocab.						.97 (.330)	.11 (.912)
	Comp.							2.04 (.041)

(continued)

Criterion Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
	Variable	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
				2C				
NWF	CLS	2.99 (.003)	1.45 (.147)	.32 (.752)	.28 (.776)	.64 (.523)	. 65 (.514)	.05 (.962)
	WWR		.16 (.876)	1.34 (.179)	.92 (.360)	1.82 (.069)	1.72 (.085)	1.19 (.234)
DIIP-S	WRC			1.60 (.109)	1.30 (.199)	2.48 (.013)	2.19 (.029)	1.71 (.087)
	WRC/min				.09 (.930)	.91 (.364)	1.05 (.293)	.31 (.759)
ORF	WRC					1.14 (.255)	1.83 (.067)	.62 (.539)
GRADE	Vocab.						.04 (.971)	1.11 (.265)
	Comp.							1.57 (.116)

(continued)

Criterion Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent Variable	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
		Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*	Z (*p)*
				2R				
NWF	CLS	**3.44 (.001)**	1.05 (.296)	.38 (.704)	.92 (.356)	1.18 (.238)	1.82 (.069)	1.00 (.316)
	WWR		.60 (.551)	1.57 (.117)	2.42 (.016)	2.66 (.008)	3.16 (.002)	2.44 (.015)
DIIP-S	WRC			.97 (.333)	2.37 (.018)	2.73 (.006)	**3.20 (.001)**	2.53 (.012)
	WRC/min				1.59 (.113)	1.54 (.124)	2.55 (.012)	1.51 (.131)
ORF	WRC					.41 (.680)	2.02 (.044)	.25 (.800)
GRADE	Vocab.						.80 (.426)	.38 (.707)
	Comp.							2.26 (.024)

(continued)

Criterion								
Measure		NWF	DIIP-S		ORF	GRADE		
	Dependent	WWR	WRC	WRC/min	WRC	Vocab.	Comp.	Total Score
	Variable	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
				S				
NWF	CLS	**3.89 (.000)**	1.00 (.317)	.34 (.737)	1.30 (.195)	1.72 (.086)	2.34 (.019)	1.66 (.098)
	WWR		.88 (.379)	1.87 (.062)	3.01 (.003)	**3.42 (.001)**	**3.86 (.000)**	**3.30 (.001)**
DIIP-S	WRC			.96 (.338)	2.81 (.005)	**3.44 (.001)**	**3.81 (.000)**	**3.39 (.001)**
	WRC/min				2.06 (.040)	2.09 (.036)	3.18 (.002)	2.28 (.022)
ORF	WRC					.67 (.505)	2.45 (.015)	.87 (.385)
GRADE	Vocab.						.80 (.424)	.19 (.846)
	Comp.							2.05 (.041)

Note. NWF = Nonsense Word Fluency; DIIP-S = Decoding Inventory for Instruction Planning – Screener; ORF= Oral Reading Fluency; CLS = Correct Letter Sounds; WWR = Whole Words Read; WRC = Words Read Correctly; Vocab. = Vocabulary; Comp. = Comprehension.

Bold: *p*≤.001

APPENDIX F

RESULTS OF MENG'S Z TEST ACROSS TEST FORMATS

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS Scores and DIBELS Next NWF CLS Scores

Test Format	5R CLS	2C CLS	2R CLS	S CLS
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS	.10 (.924)	.67 (.504)	.90 (.368)	1.97 (.049)
5R CLS		.60 (.546)	.83 (.408)	1.80 (.072)
2C CLS			.32 (.752)	1.49 (.137)
2R CLS				1.09 (.276)

Note. CLS = Correct Letter Sounds

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS Scores and DIBELS Next NWF WWR Scores

Test Format	5R CLS	2C CLS	2R CLS	S CLS
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS	.29 (.773)	.18 (.854)	.99 (.321)	1.68 (.093)
5R CLS		.44 (.662)	1.09 (.275)	1.76 (.079)
2C CLS			.81 (.420)	1.64 (.100)
2R CLS				.67 (.501)

Note. CLS = Correct Letter Sounds

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS Scores and DIIP-S WRC Scores

Test Format	5R CLS	2C CLS	2R CLS	S CLS
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS	1.13 (.259)	.40 (.690)	.46 (.649)	1.89 (.059)
5R CLS		1.47 (.143)	1.45 (.147)	**2.83 (.005)**
2C CLS			.11 (914)	1.65 (.098)
2R CLS				1.50 (.134)

Note. CLS= Correct Letter Sounds

Bold: *p* ≤ .005

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS Scores and DIIP-S WRC/min Scores

Test Format	5R CLS	2C CLS	2R CLS	S CLS
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS	.70 (.481)	.06 (.955)	.55 (.582)	1.74 (.082)
5R CLS		.77 (.442)	1.12 (.263)	2.24 (.025)
2C CLS			.48 (.629)	1.84 (.066)
2R CLS				1.24 (.215)

Note. CLS = Correct Letter Sounds

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS Scores and DIBELS Next ORF WRC Scores

Test Format	5R CLS	2C CLS	2R CLS	S CLS
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS	1.87 (.061)	.42 (.677)	.32 (.750)	1.20 (.231)
5R CLS		2.24 (.025)	2.04 (.041)	**3.01 (.003)**
2C CLS			.04 (.972)	.89 (.372)
2R CLS				.92 (.358)

Note. CLS = Correct Letter Sounds

Bold: *p* ≤ .005

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS Scores and GRADE Vocabulary Composite Scores

Test Format	5R CLS	2C CLS	2R CLS	S CLS
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS	1.43 (.153)	1.15 (.249)	.88 (.381)	.91 (.362)
5R CLS		2.35 (.019)	2.08 (.037)	2.29 (.022)
2C CLS			.11 (.914)	.15 (.885)
2R CLS				.02 (.987)

Note. CLS = Correct Letter Sounds

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS Scores and GRADE Comprehension Composite Scores

Test Format	5R CLS	2C CLS	2R CLS	S CLS
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS	1.94 (.052)	.78 (.434)	.73 (.467)	1.17 (.243)
5R CLS		2.58 (.010)	2.43 (.015)	**3.05 (.002)**
2C CLS			.06 (.954)	.50 (.617)
2R CLS				.43 (.671)

Note. CLS = Correct Letter Sounds

Bold: *p* ≤ .005

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS Scores and GRADE Total Test Scores

Test Format	5R CLS	2C CLS	2R CLS	S CLS
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS	1.91 (.057)	1.03 (.303)	.88 (.378)	1.20 (.230)
5R CLS		2.74 (.006)	2.53 (.011)	**3.04 (.002)**
2C CLS			*	.29 (.772)
2R CLS				.29 (.775)

Note. * Correlations were equal and therefore the results of Meng's Z-test was null. CLS = Correct Letter Sounds

Bold: *p* ≤ .005

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR Scores and DIBELS Next NWF CLS Scores

Test Format	5R WWR	2C WWR	2R WWR	S WWR
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR	.74 (.459)	.18 (.858)	.02 (.983)	1.61 (.106)
5R WWR		.58 (.559)	.77 (.444)	.52 (.604)
2C WWR			.16 (.876)	1.49 (.137)
2R WWR				1.40 (.162)

Note. WWR = Whole Words Read

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR Scores and DIBELS Next NWF WWR Scores

Test Format	5R WWR	2C WWR	2R WWR	S WWR
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR	.92 (.357)	.25 (.800)	.31 (.760)	1.61 (.106)
5R WWR		.70 (.486)	.72 (.471)	.35 (.729)
2C WWR			*	1.39 (.164)
2R WWR				1.15 (.249)

Note. * Correlations were equal and therefore the results of Meng's Z-test was null.

WWR = Whole Words Read

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR Scores and DIIP-S WRC Scores

Test Format	5R WWR	2C WWR	2R WWR	S WWR
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR	.75 (.455)	.04 (.965)	.13 (.895)	1.53 (.126)
5R WWR		.72 (.471)	.68 (.497)	1.96 (.050)
2C WWR			.06 (.950)	1.68 (.093)
2R WWR				1.47 (.143)

Note. WWR = Whole Words Read

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR Scores and DIBELS Next NWF DIIP-S WRC/min Scores

Test Format	5R WWR	2C WWR	2R WWR	S WWR
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR	.20 (.840)	.46 (.645)	.13 (.897)	1.32 (.188)
5R WWR		.24 (.811)	.11 (.915)	1.23 (.218)
2C WWR			.34 (.734)	2.00 (.046)
2R WWR				1.27 (.204)

Note. WWR = Whole Words Read

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR Scores and DIBELS Next ORF WRC Scores

Test Format	5R WWR	2C WWR	2R WWR	S WWR
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR	1.73 (.084)	.19 (.848)	.38 (.703)	.98 (.328)
5R WWR		1.58 (.114)	1.51 (.130)	2.50 (.013)
2C WWR			.12 (.907)	1.29 (.198)
2R WWR				1.19 (.233)

Note. WWR = Whole Words Read

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR Scores and GRADE Vocabulary Composite Scores

Test Format	5R WWR	2C WWR	2R WWR	S WWR
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR	.88 (.378)	.42 (.677)	.65 (.519)	.83 (.404)
5R WWR		1.31 (.191)	1.47 (.143)	1.54 (.124)
2C WWR			.11 (.911)	.35 (.727)
2R WWR				.17 (.869)

Note. WWR = Whole Words Read

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR Scores and GRADE Comprehension Composite Scores

Test Format	5R WWR	2C WWR	2R WWR	S WWR
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR	2.00 (.045)	.13 (.898)	.14 (.891)	.89 (.375)
5R WWR		1.93 (.054)	2.23 (.025)	2.70 (.007)
2C WWR			.23 (.817)	1.11 (.267)
2R WWR				.66 (.510)

Note. WWR = Whole Words Read

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR Scores and GRADE Total Test Scores

Test Format	5R WWR	2C WWR	2R WWR	S WWR
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR	1.73 (.085)	.09 (.931)	.38 (.702)	.97 (.334)
5R WWR		1.85 (.064)	2.14 (.032)	2.48 (.013)
2C WWR			.22 (.826)	.92 (.360)
2R WWR				.51 (.608)

Note. WWR = Whole Words Read

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS/min Scores and DIBELS Next NWF CLS Scores

Test Format	5R CLS/min	2C CLS/min	2R CLS/min	S CLS/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS/min	.87 (.387)	1.02 (.307)	1.16 (.247)	.06 (.956)
5R CLS/min		1.85 (.064)	.05 (.959)	.77 (.444)
2C CLS/min			1.97 (.049)	.94 (.347)
2R CLS/min				.98 (.327)

Note. CLS/min = Correct Letter Sounds per minute

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS/min Scores and DIBELS Next NWF WWR Scores

Test Format	5R CLS/min	2C CLS/min	2R CLS/min	S CLS/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS/min	.85 (.398)	.77 (.442)	1.55 (.121)	.89 (.376)
5R CLS/min		1.58 (.113)	.40 (.691)	.17 (.867)
2C CLS/min			2.03 (.043)	1.77 (.077)
2R CLS/min				.18 (.857)

Note. CLS/min = Correct Letter Sounds per minute

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS/min Scores and DIIP-S WRC Scores

Test Format	5R CLS/min	2C CLS/min	2R CLS/min	S CLS/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS/min	1.73 (.084)	.69 (.489)	1.69 (.090)	1.51 (.130)
5R CLS/min		1.00 (.320)	1.73 (.084)	1.73 (.084)
2C CLS/min			.61 (.540)	1.06 (.289)
2R CLS/min				.49 (.622)

Note. CLS/min = Correct Letter Sounds per minute

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS/min Scores and DIIP-S WRC/min Scores

Test Format	5R CLS/min	2C CLS/min	2R CLS/min	S CLS/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS/min	.38 (.702)	1.01 (.312)	.63 (.530)	.87 (.385)
5R CLS/min		.64 (.526)	.92 (.356)	.53 (.594)
2C CLS/min			1.54 (.123)	*
2R CLS/min				1.54 (.125)

Note. CLS/min = Correct Letter Sounds per minute; * correlations were equal and therefore the result of Meng's Z-test was null

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS/min Scores and DIBELS Next ORF WRC Scores

Test Format	5R CLS/min	2C CLS/min	2R CLS/min	S CLS/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS/min	.85 (.393)	1.37 (.170)	.85 (.395)	.08 (.940)
5R CLS/min		2.19 (.029)	1.59 (.111)	.75 (.453)
2C CLS/min			.75 (.455)	1.26 (.207)
2R CLS/min				.57 (.570)

Note. CLS/min = Correct Letter Sounds per minute

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS/min Scores and GRADE Vocabulary Composite Scores

Test Format	5R CLS/min	2C CLS/min	2R CLS/min	S CLS/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS/min	.41 (.680)	.57 (.567)	1.40 (.161)	.97 (.333)
5R CLS/min		.97 (.335)	1.57 (.116)	1.26 (.207)
2C CLS/min			.49 (.622)	.57 (.569)
2R CLS/min				.10 (.917)

Note. CLS/min = Correct Letter Sounds per minute

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS/min Scores and GRADE Comprehension Composite Scores

Test Format	5R CLS/min	2C CLS/min	2R CLS/min	S CLS/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS/min	1.53 (.127)	1.52 (.128)	.50 (.615)	.81 (.417)
5R CLS/min		**2.98 (.003)**	2.01 (.044)	.42 (.677)
2C CLS/min			1.17 (.240)	2.45 (.014)
2R CLS/min				1.38 (.167)

Note. CLS/min = Correct Letter Sounds per minute

Bold: *p* ≤ .005

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format CLS/min Scores and GRADE Total Test Scores

Test Format	5R CLS/min	2C CLS/min	2R CLS/min	S CLS/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C CLS/min	1.19 (.233)	1.28 (.200)	1.08 (.279)	1.88 (.060)
5R CLS/min		2.42 (.015)	2.14 (.033)	.90 (.368)
2C CLS/min			.48 (.634)	1.31 (.189)
2R CLS/min				.89 (.371)

Note. CLS/min = Correct Letter Sounds per minute

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR/min Scores and DIBELS Next NWF CLS Scores

Test Format	5R WWR/min	2C WWR/min	2R WWR/min	S WWR/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR/min	.84 (.401)	1.15 (.252)	.22 (.827)	.12 (.908)
5R WWR/min		1.87 (.062)	.69 (.493)	.62 (.539)
2C WWR/min			1.34 (.179)	1.29 (.197)
2R WWR/min				.04 (.970)

Note. WWR/min = Whole Words Read per minute

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR/min Scores and DIBELS Next NWF WWR Scores

Test Format	5R WWR/min	2C WWR/min	2R WWR/min	S WWR/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR/min	1.36 (.173)	.55 (.584)	1.36 (.175)	1.19 (.234)
5R WWR/min		1.78 (.074)	.27 (.791)	.03 (.979)
2C WWR/min			1.64 (.100)	1.85 (.064)
2R WWR/min				.20 (.840)

Note. WWR/min = Whole Words Read per minute

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR/min Scores and DIIP-S WRC Scores

Test Format	5R WWR/min	2C WWR/min	2R WWR/min	S WWR/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR/min	.40 (.693)	.51 (.613)	1.63 (.103)	1.08 (.282)
5R WWR/min		.84 (.398)	1.77 (.077)	1.40 (.162)
2C WWR/min			.77 (.441)	.67 (.506)
2R WWR/min				.09 (.929)

Note. WWR/min = Whole Words Read per minute

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR/min Scores and DIIP-S WRC/min Scores

Test Format	5R WWR/min	2C WWR/min	2R WWR/min	S WWR/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR/min	.41 (.683)	1.40 (.161)	.03 (.977)	.10 (.919)
5R WWR/min		.97 (.334)	.40 (.691)	.25 (.800)
2C WWR/min			1.41 (.158)	1.31 (.191)
2R WWR/min				.08 (.937)

Note. WWR/min = Whole Words Read per minute

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR/min Scores and DIBELS Next ORF WRC Scores

Test Format	5R WWR/min	2C WWR/min	2R WWR/min	S WWR/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR/min	1.71 (.087)	.50 (.620)	.05 (.961)	.52 (.603)
5R WWR/min		2.03 (.042)	1.81 (.070)	.96 (.338)
2C WWR/min			.47 (.639)	1.08 (.282)
2R WWR/min				.54 (.588)

Note. WWR/min = Whole Words Read per minute

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR/min Scores and GRADE Vocabulary Composite Scores

Test Format	5R WWR/min	2C WWR/min	2R WWR/min	S WWR/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR/min	.81 (.416)	.39 (.694)	1.15 (.252)	.06 (.951)
5R WWR/min		1.11 (.266)	1.79 (.074)	.63 (.526)
2C WWR/min			.49 (.626)	.48 (.640)
2R WWR/min				.84 (.401)

Note. WWR/min = Whole Words Read per minute

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR/min Scores and GRADE Comprehension Composite Scores

Test Format	5R WWR/min	2C WWR/min	2R WWR/min	S WWR/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR/min	2.28 (.023)	.36 (.720)	.33 (.741)	1.10 (.271)
5R WWR/min		2.41 (.016)	2.08 (.037)	.86 (.387)
2C WWR/min			.62 (.532)	1.58 (.113)
2R WWR/min				.85 (.395)

Note. WWR/min = Whole Words Read per minute

Z Scores and Significance Values from Meng's Z Test Comparing Correlations between Test Format WWR/min Scores and GRADE Total Test Scores

Test Format	5R WWR/min	2C WWR/min	2R WWR/min	S WWR/min
	Z (*p*)	Z (*p*)	Z (*p*)	Z (*p*)
5C WWR/min	1.90 (.057)	.42 (.673)	.34 (.738)	.77 (.441)
5R WWR/min		2.13 (.033)	2.25 (.025)	.87 (.383)
2C WWR/min			.17 (.864)	1.28 (.200)
2R WWR/min				.98 (.326)

Note. WWR/min = Whole Words Read per minute

BIBLIOGRAPHY

Adams, M.J. (1990). *Beginning to read.* Boston, MA: Massachusetts Institute of Technology.

American Educational Research Association. (2014). *Standards for Educational and Psychological Testing.* Washington D.C.: American Educational Research Association.

Beishuizen, J. (2008). Does a community of learners foster self-regulated learning? *Technology, Pedagogy and Education, 17(3), 183-193.*

Bell, S.M., McCallum,S., & Cox, E.A. (2003). Toward a research-based assessment of Dyslexia. *Journal of Learning Disabilities, 36(6),* 505-516.

Black, G.L. (2014). Say cheese: A snapshot of elementary teachers' engagement and motivation for classroom assessment. *Action in Teacher Education, 36,* 377-388.

Black, P., Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21(1),* 5-31.

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. Journal of Educational Computing Research, 31(1), 51-60.

Box, C., Skoog, G., & Dabbs, J.M. (2015). A case study of teacher personal practice assessment theories and complexities of implementing formative assessment. *American Educational Research Journal, 52(5),* 956-983.

Brown, J., Hinze, S., & Pellegrino, J. W. (2008). Technology and formative assessment. In T. Good (Ed.), 21st Century Education. Vol 2. Technology (pp. 245-255). Thousand Oaks, CA:Sage.

Buros Center for Testing. (2016). The mental measurements yearbook. Retrieved from http://buros.org/mental-measurements-yearbook.

Buros, O. K. (Ed.). (1938). *The 1938 mental measurements yearbook.* Highland Park, NJ: The Gryphon Press.

Buros, O. K. (Ed.). (1941). *The 1940 mental measurements yearbook.* Highland Park, NJ: The Gryphon Press.

Buros, O. K. (Ed.). (1949). *The third mental measurements yearbook.* Highland Park, NJ: The Gryphon Press.

Buros, O. K. (Ed.). (1953). *The fourth mental measurements yearbook.* Highland Park, NJ: The Gryphon Press.

Buros, O. K. (Ed.). (1959). *The fifth mental measurements yearbook.* Highland Park, NJ: The Gryphon Press.

Buros, O. K. (Ed.). (1965). *The sixth mental measurements yearbook.* Highland Park, NJ: The Gryphon Press.

Buros, O. K. (Ed.). (1972). *The seventh mental measurements yearbook.* Highland Park, NJ: The Gryphon Press.

Buros, O. K. (Ed.). (1978). *The eighth mental measurements yearbook.* Highland Park, NJ: The Gryphon Press.

Carlson, J.F., Geisinger, K.F., & Jonson, J.L. (Eds.). (2014). *The nineteenth mental measurements yearbook.* Lincoln, NE: Buros Center for Testing.

Carlson, J.F., Geisinger, K.F., & Jonson, J.L. (Eds.). (2016). *The twentieth mental measurements yearbook.* Lincoln, NE: Buros Center for Testing.

Carson, K., Gillon, G., & Boustead, T. (2011). Computer-administrated versus paper-based assessment of school-entry phonological awareness ability. *Asia Pacific Journal of Speech, Language, and Hearing, 14(2),* 85-101.

Carver, R.P. (1998). Predicting reading level in grades 1 to 6 from listening level and decoding level: Testing theory relevant to the simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 10,* 121-154.

Carver, R.P. (2003). The highly lawful relationships among pseudoword decoding, word identification, spelling, listening, and reading. *Scientific Studies of Reading, 7(2),* 127-154.

Chace, K.H., Rayner, K., & Well, A.D. (2005). Eye movements and phonological parafoveal preview: Effects of reading skill. *Canadian Journal of Experimental Psychology, 59(3),* 209-217.

Chall, J.S. (1996). *Stages of Reading Development, 2nd Ed.* Fort Worth TX: Harcourt Brace & Company.

Chen, V., & Savage, R.S. (2014). Evidence for a simplicity principle: Teaching common complex grapheme-to-phonemes improves reading and motivation in at-risk readers. *Journal of Research in Reading, 37(2),* 196-214.

Christo, C., & Davis, J. (2008). Rapid naming and phonological processing as predictors of reading and spelling. *The California School Psychologist, 13,* 7-18.

Ciampa, K. (2016). Motivating grade 1 children to read: Exploring the role of choice, curiosity, and challenge in mobile ebooks. *Reading Psychology, 37(5),* 665-705.

Coltheart, M. (2005). Modeling Reading: The Dual-Route Approach. In M. J. Snowling, C. Hulme, M. J. Snowling, C. Hulme (Eds.) , *The science of reading: A handbook* (pp. 6-23). Malden: Blackwell Publishing.

Conoley, J.C., & Kramer, J.J. (Eds.). (1989). *The tenth mental measurements yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Conoley, J.C., & Impara, J.C. (Eds.). (1995). *The twelfth mental measurements yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

De Naeghel, J., Van Keer, H.,Vansteenkiste, M., & Rosseel, Y. (2012). The relation between elementary students' recreational and academic reading motivation, reading frequency, engagement, and comprehension: A self-determination theory perspective. *Journal of Educational Psychology, 104(4)*, 1006-1021.

Deno, S.L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219-232.

Deno, S.L., & Mirkin, P. (1977). *Data-based program modification: A manual.* Reston, VA: Council for Exceptional Children.

Dynamic Measurement Group Inc. (2010). *DIBELS Next Benchmark Goals and Composite Score.*

Ehri, L.C. (1980). The development of orthographic images. In U. Frith (Ed.) *Cognitive Processes in Spelling* (pp. 311-338). London, UK: Academic Press.

Ehri, L.C. (1997). Sight word learning in normal readers and dyslexics. In B. Blachman (Ed.), *Foundations of Reading Acquisition and Dyslexia: Implications of Early Intervention* (pp. 163-190). Mahwah, NJ: Lawrence Erlbaum.

Ehri, L.C. (1998). Grapheme-phoneme knowledge is essential for learning to read words in English. In J.L. Metsala & L.C. Ehri (Eds.), *Word Recognition in Beginning Literacy* (pp. 3-40). Mahwah, NJ: Lawrence Erlbaum.

Ehri, L.C., & Wilce, L.S. (1987). Cipher versus cue reading: An experiment in decoding acquisition. *Journal of Educational Psychology, 79,* 3-13.

Every Student Succeeds Act (ESSA) of 2015. Pub. L. No. 114-95.

Fletcher, J.M., Lyon, G.R., Fuchs, L.S., & Barnes, M.A. (2007). *Learning Disabilities: From Identification to Intervention.* New York NY: The Guilford Press.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of verbal learning and verbal behavior*, *12*(6), 627-635.

Frömer, R., Dimigen, O., Niefind, F., Krause, N., Kliegl, R., & Sommer, W. (2015). Are individual differences in reading speed related to extrafoveal visual acuity and crowding? *PLOS One, 10(3),* 1-18.

Fuchs, L.S., Fuchs, D., Hosp, M.K., & Jenkins, J.R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5(3),* 239-256.

Gardner, D. P. (1983). A nation at risk. *Washington, DC: The National Commission on Excellence in Education, US Department of Education*.

Geisinger, K.F., Spies, R.A., Carlson, J.F., & Plake, B.S. (Eds.). (2007). *The seventeenth mental measurements yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Gentry, J.R. (2006). *Breaking the code: The new science of reading and writing.* Portsmouth, NH: Heinemann.

Good III, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, *5*(3), 257-288.

Good III, R.H., & Kaminski, R.A. (2011). *DIBELS Next Assessment Manual.* Dynamic Measurement Group.

Good III, R.H., & Kaminski, R.A. (2010). *DIBELS Next Nonsense Word Fluency.* Dynamic Measurement Group.

Good III, R.H., & Kaminski, R.A. (2010). *DIBELS Next Oral Reading Fluency.* Dynamic Measurement Group.

Gottardo, A., Chiappe, P., Siegel, L.S., & Stanovich, K.E. (1999). Patterns of word and nonword processing in skilled and less-skilled readers. *Reading and Writing: An Interdisciplinary Journal, 11,* 465-487.

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

Gough, P.B. (1996). How children learn to read and why they fail. *Annals of Dyslexia, 46*, 3-20.

Gray, L., Thomas, N., & Lewis, L. (2010). Teachers’ Use of Educational Technology in U. S., Public Schools: 2009 (NCES 2010-040). National Center for Education Statistics, Institute of Education Sciences, U. S. Department of Education. Washington, DC.

Hernandez, D. J. (2011). *Double jeopardy: How third-grade reading skills and poverty influence high school graduation.* New York: Annie E. Casey Foundation.

Hosp, J.L., & Ardoin, S.P. (2008). Assessment for instructional planning. *Assessment for Effective Intervention, 33(2),* 69-77.

Hosp, M.K., Hosp, J.L., & Howell, K.W. (2016). *The ABCs of CBM* (2nd ed.). New York, NY: Guilford Press.

Hosp, M.K. (2016). *Decoding Inventory for Instructional Planning.*

Impara, J.C., & Plake, B.S. (Eds.). (1998). *The thirteenth mental measurements yearbook.* Lincoln, NE: Buros Institute for Mental Measurements.

Jenkins, J.R., Fuchs, L.S., van den Broek, P., Espin, C., & Deno, S.L. (2003). Accuracy and fluency in list and context reading of skilled and RD groups: Absolute and relative performance levels. *Learning Disabilities Research and Practice, 18(4),* 237-245.

Jenkins, J.R., Fuchs, L.S., van den Broek, P., Espin, C., & Deno, S.L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Pscyhology, 94(4)*, 719-729.

Jenkins, J.R., & Fuchs, L.S. (2012). Curriculum-based measurement: The paradigm, history, and legacy. In Espin, C.A., McMaster, K.L., Rose, S., & Wayman, M.M. (Eds.), *A measure of success: The influence of curriculum-based measurement on education* (pp.7-23). Minneapolis, MN: University of Minnesota Press.

Jones, F.G., Gifford, D., Yovanoff, P., Al Otaiba, S., Levy, D., & Allor, J. (in press). Alternate assessment formats for progress monitoring students with intellectual disabilities an below average IQ: An exploratory study. *Focus on Autism.*

Jones, M.W., Branigan, H.P., & Kelly, M.L. (2009). Dyslexic and nondyslexic reading fluency: Rapid auomatized naming and the important of continuous lists. *Psychonomic Bulletin and Review, 16(3),* 567-572.

Joshi, R.M., & Aaron, P.G. (2002). Naming speed and word familiarity as confounding factors in decoding. *Journal of Research in Reading, 25(2),* 160-171

Knutson, J.S., Simmons, D.C., Good, R., & McDonagh, S.H. (2004). Specially designed assessment and instruction for children who have no responded adequately to reading intervention. *Assessment for Effective Intervention, 29(4),* 47-58.

Kramer, J.J., & Conoley, J.C. (Eds.). (1992). *The eleventh mental measurements yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the no child left behind act of 2001. *Educational Researcher*, *31*(6), 3-16.

Marx, C., Hutzler, F., Schuster, S., & Hawelka, S. (2016). On the development of parafoveal preprocessing: Evidence from the incremental boundary paradigm. *Frontiers in Psychology, 7,* 1-13.

McCardle, P., & Chhabra, V. (2004). *The Voice of Evidence in Reading Research*: Brookes Publishing Company.

Meng, X. L., Rosenthal, R., & Rubin, D. (1992). Comparing correlated correlation coeffecients. Psychological Bulletin, 111, 172-175.

Mitchell, J.V. (Ed.). (1985). *The ninth mental measurements yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Moats, L, & Tolman, C. (2009). *Language Essentials for Teachers of Reading and Spelling (LETRS): The Challenge of Learning to Read (Module 1*). Boston, MA: Sopris West.

Morgan, P.L., & Fuchs, D. (2007). Is there a bidirectional relationship between children's reading skills and reading motivation? *Exceptional Children, 73(2),* 165-183.

Murray, M.S., Munger, K.A., & Clonan, S.M. (2012). Assessment as a strategy to increase oral reading fluency. *Intervention in School and Clinic, 47(3),* 144-151

National Center for Educational Statistics. (2015). The NAEP reading achievement levels by grade. Retrieved from https://nces.ed.gov/nationsreportcard/reading/achieve.aspx.

National Center for Education Statistics. (2015). *The nation's report card: Mathematics and reading assessments*. Washington, DC: Institute of Educational Sciences: U.S. Department of Education.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Washington, DC: Authors.

National Reading Panel (US), National Institute of Child Health, & Human Development (US). (2000). *Report of the national reading panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health and Human Development, National Institutes of Health.

Nelson, J. S., Jayanthi, M., Epstein, M. H., & Bursuck, W. D. (2000). Student preferences for adaptations in classroom testing. *Remedial and Special Education*, *21*(1), 41-52.

No Child Left Behind (NCLB) Act of 2002, Pub. L. No. 107-110.

Oosterhof, A. (2003). *Developing and using classroom assessments.* Upper Saddle River, NJ: Merrill Prentice Hall.

Pearson, P. D. (2004). The Reading Wars. *Educational Policy*, *18*(1), 216–252.

Perfetti, C.A. (1984). Reading acquisition and beyond: Decoding includes cognition. *American Journal of Education, 93*, 40-60.

Perfetti, C.A. (1986). Continuities in reading acquisition, reading skill, and reading disability. *Remedial and Special Education*, *7,* 11-21.

Pierce, M., Katzir, T., Wolf, M., & Noam, G. (2010). Examining the construct of reading among dysfluent urban children: A factor analysis approach. *Journal of Literacy Research*, *42,* 124-158.

Plake, B.S., & Impara J.C. (Eds.). (2001). *The fourteenth mental measurements yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Plake, B.S., Impara, J.C., & Spies, R.A. (Eds.). (2003). *The fifteenth mental measurements yearbook.* Lincoln, NE: Bros Institute of Mental Measurements.

Polloway, E. A., Bursuck, W. D., Jayanthi, M., Epstein, M. H., & Nelson, J. S. (1996). Treatment acceptability: Determining appropriate interventions within inclusive classrooms. *Intervention in School and Clinic*, *31*(3), 133-144.

Pomplun, M., Ritchie, T., & Custer, M. (2006). Factors in paper-and-pencil and computer reading score differences at the primary grades. *Educational Assessment*, *11(2),* 127-143.

Pritchard, S. C., Coltheart, M., Palethorpe, S., & Castles, A. (2012). Nonword reading: Comparing dual-route cascaded and connectionist dual-process models with human data. *Journal Of Experimental Psychology: Human Perception And Performance*, *38*(5), 1268-1288.

Quellmalz, E. S. & Pellegrino, J. W. (2009). Technology and testing. Science, 323, 75-78.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

Randi, J., & Corno, L. (2000). Teacher innovations in self-regulated learning. In M. Boeaerts, P. R. Pintrich, & M. Zeidner (Eds.), Handbook of self-regulation (pp. 651–685). San Diego, CA: Academic Press.

Rathvon, N. (2004). *Early reading assessment.* New York, NY: Guilford Press.

Reutzel, D.R., Brandt, L., Fawson, P.C., & Jones, C.D. (2014). Exploration of the consortium on reading excellence phonics survey: An instrument for assessing primary-grade student' phonics knowledge. *The Elementary School Journal, 115(1),* 49-72.

Ritchey, K.D. (2008). Assessing letter sound knowledge: A comparison of letter sound fluency and nonsense word fluency. *Exceptional Children, 74(4),* 487-506.

Robbins, K.P., Hosp, J.L., Hosp, M.K., & Flynn, L.J. (2010). Assessing specific grapho-phonemic skills in elementary students. *Assessment for Effective Intervention, 36,* 21-34.

Roman, S. P. (2004). Illiteracy and older adults: Individual and societal implications. *Educational Gerontology, 30*(2), 79-93.

Salvia, J., & Ysseldyke, J.E. (2001). *Assessment* (8th ed.). Boston: Houghton Mifflin.

Salvia, J., Ysseldyke, J.E., & Bolt, S. (2013). *Assessment in special and inclusive education.* Belmont, CA: Cengage Learning.

Shaywitz, S.E. (2003). *Overcoming dyslexia: a new and complete science-based program for reading problems at any level.* New York, NY: Alfred A. Knopf.

Shepard, L., Hammerness, K., Darling-Hammond, L., Rust, F. (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 275-326). San Francisco, CA: Jossey-Bass.

Spies, R.A., & Plake, B.S. (Eds.). (2005). *The sixteenth mental measurements yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Spies, R.A., Carlons, J.F., & Geisinger, K.F. (Eds.). (2010). *The eighteenth mental measurements yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Stanovich, K.E. (1981). Relationships between word decoding speed, general name-retrieval ability, and reading progress in first-grade children. *Journal of Educational Psychology, 73(6),* 809-815.

Stanovich, K. E. (1991). Word recognition: Changing perspectives. *Handbook of reading research*, *2*, 418-452.

Stiggins, R.J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, 758-765.

Tobin, R. (2008). Strategies for differentiating in the language arts: Positioning struggling students for early success. *English Quarterly Canada, 38(2/3*), 57–65.

Torgesen, J.K. (1998). Catch them before they fall: Identification and assessment to prevent reading failure in young children. *American Educator,* 32-39.

Torgesen, J.K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disablities Research and Practice, 15(1),* 55-64.

Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of school psychology*, *40*(1), 7-26.

Vinovskis, M.A. (1998). Overseeing the nation's report card. (Report to the National Assessment Governing Board). U.S. Department of Education.

Weiser, B.L. (2012). Ameliorating reading disabilities early: examining an effective encoding and decoding prevention instruction model. *Learning Disability Quarterly, 36(3),* 161-177.

Whitaker, C. P., Gambrell, L. B., & Morrow, L. M. (2004). Reading comprehension instruction for all students. In E. Silliman & L. C. Wilkinson (Eds.), *Language and Literacy Learning in Schools* (pp. 130-150). New York The Guilford Press.

Wilkins, A., Cleave, R., Grayson, N., & Wilson, L. (2009). Typography for children may be inappropriately designed. *Journal of Research in Reading, 42(4),* 402-412.

Williams, K.T. (2001). *The Group Reading Assessment and Diagnostic Evaluation.* San Antonio, TX: Pearson.

Yell, M. L. (2016). *The law and special education (4th ed.).* Boston, MA: Pearson.