AN ARGUMENT-BASED APPROACH TO EARLY LITERACY CURRICULUM-BASED MEASURE VALIDATION WITHIN MULTI-TIERED SYSTEMS OF SUPPORT IN READING: DOES INSTRUCTIONAL EFFECTIVENESS MATTER?

by

MARISSA PILGER SUHR

A DISSERTATION

Presented to the Department of Special Education and Clinical Sciences
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2021

DISSERTATION APPROVAL PAGE

Student: Marissa Pilger Suhr

Title: An Argument-Based Approach to Early Literacy Curriculum-Based Measure Validation Within Multi-Tiered Systems of Support in Reading: Does Instructional Effectiveness Matter?

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Special Education and Clinical Sciences by:

Hank Fien	Chair
Gina Biancarosa	Core Member
Benjamin Clarke	Core Member
Nancy Nelson	Core Member
Elizabeth Budd	Institutional Representative

and

Kate Mondloch	Interim Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2021.

DISSERTATION ABSTRACT

Marissa Pilger Suhr

Doctor of Philosophy

Department of Special Education and Clinical Sciences

June 2021

Title: An Argument-Based Approach to Early Literacy Curriculum-Based Measure Validation Within Multi-Tiered Systems of Support in Reading: Does Instructional Effectiveness Matter?

Early literacy curriculum-based measures (CBMs) are widely used as universal screeners within multi-tiered systems of support in reading (MTSS-R) for (1) evaluating the overall effectiveness of the reading system and (2) assigning students to supplemental and intensive interventions. Evidence supporting CBM validity for these purposes have primarily relied on diagnostic accuracy statistics obtained from evaluations of CBMs' discriminative (i.e., sensitivity and specificity) and predictive (i.e., likelihood ratios, posttest probabilities) ability across various lag times and instructional contexts. The treatment paradox has been identified as a potential source of bias which may systematically alter diagnostic accuracy statistics when there is substantial lag time between administrations of the screener and outcome measure within medical diagnostic accuracy studies, particularly for conditions that lie on a continuum such as reading difficulties. However, the impact of the treatment paradox on early literacy screener diagnostic accuracy statistics in the context of MTSS-R is unknown.

The current study examines the degree to which the treatment paradox, in the form of reading instruction, alters the diagnostic accuracy of a nonsense word fluency screener across different lag times. Concurrent and predictive validity coefficients and

diagnostic accuracy statistics are examined within the context of a randomized controlled trial for meaningful differences across time points, lag times and levels of instructional effectiveness across two different outcome measures.

CURRICULUM VITAE

NAME OF AUTHOR: Marissa Pilger Suhr

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Williams College, Williamstown, Massachusetts

DEGREES AWARDED:

Doctor of Philosophy, School Psychology, 2021, University of Oregon
Master of Science, Special Education, 2019, University of Oregon
Bachelor of Arts, Psychology, 2011, Williams College

AREAS OF SPECIAL INTEREST:

Multi-Tiered Systems of Supports in Reading
Data-Based Decision-Making
Teacher Professional Development and Coaching

PROFESSIONAL EXPERIENCE:

School Psychologist Intern, Springfield Schools, 2020 to present

Research Assistant, Center on Teaching and Learning, University of Oregon, 2019 to present

Research and Technical Assistance Graduate Employee, Center on Teaching and Learning, University of Oregon, 2018 to 2019

Research Assistant Graduate Employee, Behavioral Research and Teaching, University of Oregon, 2015 to 2018

Literacy Tutoring Site Coordinator, Reading Partners, 2013 to 2015

Project Coordinator, Hinshaw ADHD Lab, University of California Berkeley, 2012

Special Education Instructional Assistant, Raskob Day School, 2011 to 2012

GRANTS, AWARDS, AND HONORS:

Dynamic Measurement Group Award, University of Oregon, 2018

Dynamic Measurement Group Award, University of Oregon, 2017

Council for Learning Disabilities 1st Annual Leadership Institute, 2017

Dynamic Measurement Group Award, University of Oregon, 2016

PUBLICATIONS:

Pilger Suhr, M., Nese, J. F. T., & Alonzo, J. (2021). Parallel Reading and Mathematics Growth for English Learners: Does Timing of Reclassification Matter? *Journal of School Psychology, 85,* 94-112.

Clarke, B. S., Doabler, C. T., Sutherland, M., Suhr, M. P., Kiru, E. W. (in press). Intensifying early numeracy interventions. In D. P. Bryant (Ed.), *Intensifying Mathematics Interventions for Struggling Students,* Guilford Press.

Fien, H., Nelson, N. J., Smolkowski, K., Kosty, D., Pilger, M., Baker, S. K., Smith, J. L. M. (2020). A Conceptual Replication Study of the Enhanced Core Reading Instruction MTSS-Reading Model. *Exceptional Children.* Advance online publication. https://doi.org/10.1177/0014402920953763

Shanley, L., Strand Cary, M., Turtura, J., Clarke, B., Sutherland, M., & Pilger, M. (2019). Individualized instructional delivery options: Adapting technology-based interventions for students with attention difficulties. *Journal of Special Education Technology, 35*(3), 119-132. https://doi.org/10.1177/0162643419852929

ACKNOWLEDGMENTS

I wish to thank Dr. Hank Fien for his assistance in the preparation of this manuscript, as well as Drs. Gina Biancarosa, Ben Clarke, Nancy Nelson, and Elizabeth Budd for their thoughtful insights and ongoing support throughout the dissertation process. I thank my colleagues and friends in the School Psychology program who have always been there to collaborate, innovate, and commiserate throughout my time at the University of Oregon. I thank my family for helping me believe that I could make it through a Ph.D. program. And last but not least I thank Julian for his delicious meals, superb household management, and listening ear, which have allowed me to get to where I am today. The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A090104 to the University of Oregon.

TABLE OF CONTENTS

Chapter Page

Chapter Page

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

Table Page

I. INTRODUCTION

Despite a wealth of reading research (e.g. National Reading Panel, 2000) and substantial federal funding and efforts dedicated to improving students' literacy skills over the past several decades (e.g. Reading First, Every Student Succeeds Act), approximately two thirds of 4th grade students in the United States continue to perform below proficient on the National Assessment of Education Progress (NAEP, 2019). In response to persistent concerns regarding students' reading proficiency, the 2004 reauthorization of the Individuals with Disabilities in Education Act (IDEA) enacted legislation enabling schools to dedicate special education funds to the provision of intervention services to students at risk for reading difficulties (IDEA, 2004).

Since this reauthorization, schools have increasingly begun to use response to intervention (RTI) or multi-tiered systems of support in reading (MTSS-R) as comprehensive frameworks for seamlessly integrating evidence-based instructional strategies to improve students' reading outcomes (Balu, 2015; Gersten et al., 2009; Samuels, 2011). MTSS-R is a tiered service delivery model intended to improve reading outcomes for all students within a school by providing increasingly intensive and individualized evidence-based instruction to students based on level of need. The theoretical foundation of MTSS-R is based on decades of research that highlights the need for prevention and early reading intervention to address reading problems in kindergarten and first grade before reading trajectories become established and increasingly difficult to alter (Juel, 1988; Kame'enui & Carnine, 1998; McCardle et al., 2001; Scarborough, 1998).

The Role of Universal Screening Within MTSS-R

Universal screening is arguably the cornerstone to an effective MTSS-R model (Fuchs et al., 2003; Gersten et al., 2009). Universal screening is a process by which a population of students is evaluated to identify the presence or absence of a specific condition of interest (Jenkins et al., 2007). In the context of MTSS-R, schools use universal screeners to identify the presence or absence of risk for reading difficulties. This information can help educators make instructional decisions within their schools that accomplish two primary purposes.

First, schools may use universal screening scores to evaluate the overall effectiveness of their reading system (Deno, 2003; Tindal, 1989; Tindal, 2013). Screeners assign a risk status to students (e.g. at risk, some risk, low risk), and based on the proportion of students who are classified as at risk versus not at risk, a school may decide to focus efforts on improving Tier 1 core instruction versus Tier 2 or 3 supplemental intervention. For example, if screening data suggests that 80% of students in a school are at some risk for developing reading difficulties, this is an indication that core instruction is not meeting students' needs and that school leadership should problem-solve around adjustments to instruction for all students. Conversely, if screening data indicates that only 5% of students are demonstrating reading risk, school leadership can prioritize intensifying intervention for just a few students. For this purpose, schools rely on screeners to accurately differentiate between students who actually are and are not at risk for reading difficulties to make systems-level instructional decisions with the goal of decreasing the total number of at-risk students.

Second, schools may use screening scores to assign individual students who are classified as at risk to evidence-based supplemental and intensive intervention (Silva et al., 2020). Evidence suggests that these Tier 2 and 3 supports can accelerate the rate of learning for at-risk students such that they catch up to their grade-level peers, reducing the number of students who go on to develop pervasive reading difficulties (Fuchs et al., 2012; Gersten et al., 2009; Jenkins et al., 2013; Johnson et al., 2010; Mellard et al., 2009; Wanzek, et al., 2016). For this purpose, schools rely on screeners to accurately rule in or out whether *individual students* are likely to develop reading difficulties over time so that those students who are most at risk can be provided with supplemental or intensive intervention.

Curriculum-based measures (CBMs) are one of the most widely used tools for universal screening within MTSS-R for these two primary screening purposes (Fuchs & Vaughn, 2012; Gersten et al., 2009). CBMs are fluency-based measures which assess students' skills in key reading-related areas. CBM measures of oral reading fluency were first developed in the 1970s at the University of Minnesota for instructional planning and data-based decision making for individual students within special education (Deno, 1989). In the decades since their inception, CBMs have been adopted for a wider variety of purposes and have been expanded to measure a broader range of skills (Espin et al., 2012; Shinn, 1998). Presently, CBMs are used for everything from universal screening to progress monitoring to program evaluation to special education eligibility (Fuchs & Vaughn, 2012), with subtests targeting foundational reading skills, including phonological awareness, decoding, and reading comprehension. Early literacy CBMs, which measure phonological awareness, letter-sound correspondences, and real and

nonword reading skills, have grown especially popular as screening tools in kindergarten and first grade when measures of oral reading fluency may not be sufficiently sensitive to differentiate among student performance (Catts et al., 2009; Gersten et al., 2009; Silberglitt & Hintze, 2005).

An Argument-Based Approach to Validation of Early Literacy CBMs

Early literacy CBMs have been widely touted by researchers and practitioners alike as gold standard screening tools for both evaluating overall school systems' effectiveness as well as predicting individual students' likelihood of developing future reading difficulties (e.g., Gersten et al., 2009) and many states have adopted policies requiring schools to collect this universal screening data for all students in grades K-2 (National Center on Improving Literacy [NCIL], 2020). Given their widespread use, it is important to have clear evidence of early literacy CBMs' validity for these purposes.

Kane's argument-based approach to test validation provides a helpful framework for obtaining this evidence (Kane, 1992, 2006). The argument-based approach to test validation maintains that the primary purpose of all test scores is to support broader interpretations about a student based on that student's observed performance on a test. For example, within MTSS-R educators may use a student's test score not just to describe that student's skills with the behavior sampled on the test (e.g. decoding nonwords), but also as a hypothesis about the student's skills in a broader construct of interest (e.g. reading proficiency) or the student's probability of reading success across a school year.

Based on these broad interpretations, educators use a given test score or scores to inform instructional decisions, with the nature of the interpretation dictating the kind of the decisions that are made. For example, if a student's test score is interpreted as

indicating that the student is highly likely to develop reading difficulties, that student may be immediately assigned to supplemental or intensive intervention. In contrast, if the same test score is interpreted as demonstrating only a slight increased likelihood of developing reading difficulties, the team may choose to monitor the student's progress more closely before providing them with supplemental intervention. Thus, interpretations made based on test scores have a sizable impact on the types of decisions that are made; however, these interpretations are often made subconsciously, without deliberate consideration by educators of their influence on instructional decision-making.

Kane's argument-based approach to test validation reasons that to comprehensively validate a test, the myriad ways in which a test score may be interpreted and used in a given context must be explicitly articulated and evaluated. Thus, a test must be evaluated not only on whether it reliably assesses a construct of interest, which provides evidence of the validity of a test score's *interpretations,* but also on how well it aids in instructional decision making, providing evidence of the validity of a test score's *uses.* A comprehensive evaluation of how well a test helps with instructional decision making requires an explicit consideration of the social consequences of using the test for its various purposes, a concept called consequential validity (Messick, 1975, 1989). Depending on the specific inferences made about a student's test score, the same test may result in different decisions made in different settings, and so acceptable evidence of a test's consequential validity may vary depending on the context. To comprehensively evaluate early literacy CBMs for the purposes of (a) evaluating the effectiveness of reading systems and (b) assigning students to supplemental and intensive instruction

within MTSS-R, then, requires an explicit examination of evidence supporting each of early literacy CBMs' proposed interpretations and uses.

Evaluating CBM Test-Score Interpretations

Researchers and practitioners intending to evaluate the appropriateness of a universal screener for use within their school's MTSS-R framework should consider the proposed interpretations and uses of the screener in that context and determine whether the test has empirical evidence supporting these interpretations and uses. Historically, evidence for CBM validity as a universal screener was limited to support for these tests' proposed interpretations rather than their proposed uses. This evidence was primarily obtained through evaluations of CBMs' criterion-related validity, with strong screener criterion validity providing evidence that the student's performance on the screener is indicative of that student's reading skills overall. Specifically, studies have examined screeners' concurrent and predictive validity, statistics which provide information about the degree to which the screener is related to an established measure of a broader construct of interest (e.g., reading proficiency) when the two tests are administered at the same time point (i.e., concurrent validity) and at two different time points (i.e., predictive validity) (Kilgus et al., 2014).

A large number of studies over the past several decades have provided this evidence for oral reading fluency CBMs' proposed interpretations within MTSS-R. These measures have been found to be strongly related to widely recognized measures of reading achievement, with meta-analyses reporting mean correlation coefficients between .56 and .73 for the relation between oral reading fluency and both statewide achievement tests and standardized norm-referenced assessments (Reschly et al., 2009; Yeo, 2010).

Criterion validity of early literacy CBMs such as nonsense word fluency and word reading fluency measures suggest that these measures are highly related to measures of oral reading fluency, with correlations ranging from .68-.82 for measures of nonsense word fluency (e.g., Burke & Hagan-Burke, 2007; Cummings et al., 2011; Fien et al., 2010; Harn et al., 2008; January & Klingbeil, 2020) and .80-.93 for measures of word reading fluency (Fuchs et al., 2004). These measures are also related to broader measures of reading achievement, with correlations ranging from .60-.73 for nonsense word fluency (e.g., Fien et al., 2008; Fien et al., 2010; January & Klingbeil, 2020), and .66-.79 for word reading fluency (Fuchs et al., 2004; January & Klingbeil, 2020).

These criterion-related data provide evidence that early literacy CBM test scores are indicative of students' overall reading skills, a necessary first step in interpreting data. However, criterion-related validity is not sufficient for providing evidence for early literacy CBMs' proposed *uses* within MTSS-R (Burns, 2012). That is, criterion-related validity alone does not provide evidence regarding how accurately an early literacy CBM discriminates between individuals with and without reading difficulties for the purpose of evaluating reading systems or how accurately an early literacy CBM predicts a student's likelihood of future reading difficulties for the purpose of assigning students to interventions. To validate early literacy CBMs for these uses, an evaluation of these tests' diagnostic accuracy is also necessary (Jenkins et al., 2007; Kilgus et al., 2014).

Evaluating CBM Test-Score Uses

Diagnostic accuracy evaluations have grown increasingly prevalent in the past decade for determining the validity of CBMs for predicting reading risk. The general framework for evaluating the diagnostic accuracy of screening tools originated in radio

signal detection work (Petersen et al., 1954), and has more recently been applied to screening tests across the fields of medicine (Pepe, 2003), epidemiology and public health (Fleiss, 1981), and psychology (Swets, 1996). A brief overview of key diagnostic accuracy statistics is provided below.

Diagnostic Accuracy Overview. Broadly, diagnostic accuracy evaluations are conducted to determine a test's accuracy at (1) discriminating between groups of individuals with and without a condition and (2) predicting an individual's membership in one of these two groups. The specific group, or population, an individual belongs to is determined by the individual's classification based on some gold standard outcome measure which is accepted as an individual's "true" condition (Deeks, 2001). Diagnostic accuracy evaluations determine the degree to which a screener classifies individuals as belonging to the population of individuals who have a given condition versus the population of individuals who do not have the condition in the same way that a more widely accepted test classifies individuals. To conduct a diagnostic accuracy evaluation every individual in the sample is administered both the screener and outcome measure and a prediction is made about which population each individual belongs to based on their screener score.

When evaluating the diagnostic accuracy of a screener, a key assumption is that the outcome measure accurately classifies individuals into these two populations (those that have the condition, and those that do not have the condition) (Smolkowski & Cummings, 2015). For conditions which fall on a continuous scale, such as reading difficulty, the two populations are determined by test makers or evaluators, who decide on a set cut-score on the outcome measure which artificially classifies students into one

of two groups: students whose scores fall below the cut-score, and thus have the condition (e.g., have reading difficulties), and students whose scores fall above the cut-score, and so do not have the condition (e.g., do not have reading difficulties). The screener is then evaluated for its ability to classify students into the group of students with reading difficulties or the group of students without reading difficulties in the same way that the outcome measure classifies students, and based on this evaluation test evaluators decide on an optimal cut-score on the screener that is indicative of risk for reading difficulties; students who fall above the cut-score are considered not at risk, and students who fall below the cut-score are considered at risk.

Because no screening measure can be one hundred percent accurate at classifying every single student, students will fall into one of four categories based on the screener cut-score for risk. The category a student is assigned to will depend on their performance on the screener (i.e., the observed reading behavior) and their actual level of reading difficulty (i.e., true reading skill, as measured by performance on the outcome measure of reading achievement). Of all the students who fall below the screener cut-score for risk, a proportion of students will truly have reading difficulties (True Positive [TP]), and a proportion of students will actually not have reading difficulties (False Positive [FP]). Similarly, of the students who fall above the cut-score for risk, a subset of students will truly not have reading difficulties (True Negative [TN]) and a group of students will actually have reading difficulties (False Negative [FN]).

These four categories are inextricably linked and dependent on the predetermined screener risk cut-score; as the cut-score for risk increases, the screener will accurately classify more students who truly have reading risk but will also inaccurately classify

more students as at risk who actually do not have reading difficulties. Similarly, as the cut-score for risk decreases, the screener will accurately classify more students who truly do not have reading difficulties but will also inaccurately classify more students as not at risk when they actually have reading difficulties. When choosing a cut-score for risk on the screening measure, test makers must try to create an optimal balance between these four categories and will choose to prioritize different diagnostic accuracy statistics in making their ultimate decision. These diagnostic accuracy statistics are described below.

Overall Test Accuracy. The accuracy with which a screener classifies students as either at risk or not at risk is evaluated through the use of receiver operating characteristic (ROC) curves. ROC curves describe the proportion of time a screener accurately classifies students as at risk (true positive fraction [TPF]) relative to the proportion of time a screener inaccurately classifies students as at risk (false positive fraction [FPF]) across all possible screener scores. A screener that does a good job of maximizing the TPF while minimizing the FPF will have a higher area under the curve (AUC) value, which provides a summary of the overall performance of the screener across all possible cut-score or decision thresholds. The AUC can be described as the likelihood that the screener will accurately classify a randomly chosen pair of individuals, one from the at-risk population and one from the not-at-risk population. AUC values range from .00 to 1.00; AUC values of .00 indicate that a screener would inaccurately classify students 100% of the time, AUC values of .50 indicate that a screener provides no useful information, and would accurately classify individuals 50% of the time, and AUC values of 1.0 indicate that a screener would classify individuals with 100% accuracy. Generally, screeners with AUC values of .95 and above are considered excellent, screeners with

AUC values of .85-.95 are considered very good, screeners with AUC values of .75-.85 are considered reasonable, and screeners with AUC values below .75 are considered poor and should not be used for decision making (Smolkowski & Cummings, 2015; Swets, 1988).

Sensitivity and Specificity Rates. In addition to overall accuracy statistics, ROC curves produce a number of valuable statistics for decision making which describe the diagnostic accuracy of a screener associated with specific risk cut-scores. The *sensitivity* of a measure signifies the proportion of students who were correctly identified by the screener as at risk in relation to the entire population of students who truly have reading difficulties. In other words, sensitivity is concerned with only the population of individuals with the condition, and the sensitivity value indicates how well a test can recognize an individual with the condition. When evaluating reading screeners, sensitivity refers to the population of students with reading difficulties and is calculated by dividing true positives by the sum of true positives and false negatives. High sensitivity rates are important to a reading screener, as they indicate that the screener has accurately identified most or all students who truly are at risk for reading difficulties and are in need of supplemental supports and has thus minimized the number of students who are truly at risk but were not identified. Education researchers generally agree that sensitivity rates should be prioritized within MTSS-R, and screeners should have a minimum sensitivity value of .80 to .90 to be appropriate for use in these settings (Jenkins et al., 2007; Petscher et al., 2011), where the cost of not providing intervention services to students who are at risk for reading difficulties is considered more

problematic than unnecessarily providing intervention services to students who do not need them.

Specificity refers to the proportion of students who were correctly identified by the screener as not at risk in relation to the population of students who truly do not have reading difficulties. In other words, specificity is concerned with only the population of individuals without the condition, and the specificity value indicates how well the test can recognize an individual without the condition. In regard to reading screeners, specificity refers to the subgroup of students without reading difficulties and is calculated by dividing true negatives by the sum of true negatives and false positives. High specificity rates are important to a reading screener, as they indicate that the screener is accurately identifying *only* those students who are at risk for reading difficulties, and not misclassifying as at risk students who are on track for reading success. Thus, high specificity rates should minimize the likelihood of the school system being overwhelmed with providing intervention services to students who are not actually in need of them. Minimally acceptable specificity rates have been more widely debated than sensitivity, with researchers generally promoting specificity ranging from .70 to .80 and higher (Catts et al., 2009; Compton et al., 2010).

Test makers generally rely on sensitivity and specificity values to set screener cut-scores for risk, prioritizing either sensitivity, specificity, or a balance of the two in deciding on the optimal risk cut-score. Sensitivity and specificity are the most widely used diagnostic accuracy statistics for evaluating the accuracy of a screener and setting screener risk cut-scores because they are population-based statistics, meaning that they are thought to be reasonably robust across settings. In other words, it is generally

accepted that sensitivity and specificity rates obtained from a research study will apply across diverse school settings.

Likelihood Ratios. Likelihood ratios indicate how much more likely a specific screening result is for individuals who have the condition than for individuals who do not have the condition (Choi, 1998). Likelihood ratios take into account both sensitivity and specificity in their calculations, allowing for a comparison between the population of students with true reading difficulties and the population of students who truly do not have reading difficulties. Because likelihood ratios are derived from sensitivity and specificity values, they are population-based statistics and not impacted by base rate; thus, it is generally accepted that these ratios are applicable to a variety of contexts.

Likelihood ratios are calculated by dividing the likelihood of a given test result for individuals with the condition (e.g., reading difficulties) by the likelihood of that same test result for individuals without the condition (e.g., no reading difficulties). *Positive likelihood ratios* indicate how much more likely a positive test result (e.g., classification of "at risk") is for individuals who have the condition (e.g., true reading difficulties) than for individual who do not have the condition (e.g., truly no reading difficulties). In contrast, *negative likelihood ratios* indicate how much more likely a negative test result (e.g., classification of "not at risk") is for individuals who have the condition than for individuals who do not have the condition (Kent & Hancock, 2016).

Likelihood ratios near 1 indicate that the screener does not meaningfully change the likelihood of having the condition—thus indicating that the screener has little use. Likelihood ratios greater than 1 indicate a progressively increased likelihood of the condition, while likelihood ratios smaller than 1 indicate a progressively decreased

likelihood of the condition. While the field of education has not settled on minimum acceptable likelihood ratios, in the medical world likelihood ratios of 2 to 5 are generally interpreted as small increases in likelihood of the condition, while likelihood ratios of 5 to 10 are interpreted as moderate increases, and likelihood ratios above 10 are interpreted as large increases in likelihood of the condition. In medicine, positive likelihood ratios of 10 or higher are generally considered meaningful for ruling in the presence of a condition (e.g., cancer). Conversely, likelihood ratios of 0.2 to 0.5 indicate small *decreases* in likelihood of the condition, while likelihood ratios of 0.1 to 0.2 indicate moderate decreases, and less than 0.10 indicate large and conclusive decreases in the likelihood of the condition. In medicine, negative likelihood ratios of less than .10 are considered meaningful for ruling out the presence of the condition (Grimes & Schulz, 2005; McGee, 2001).

In education, where CBMs are used to make a variety of decisions, educators should consider the stakes of the decision being made when determining an acceptable likelihood ratio for their purposes. For example, a likelihood ratio closer to 1 may be acceptable for making a decision about whether to assign a student to a brief supplemental intervention, while a likelihood ratio much further from 1 would be necessary for making decisions about whether a child's level of risk warrants fast-tracking to more intensive instructional supports.

Posttest Probabilities. Likelihood ratios can also be used to calculate the posttest probability of having a condition given a specific screening result (VanDerHeyden, 2011, 2013). In other words, researchers and educators can use likelihood ratios to help determine the probability of an individual student actually having reading difficulties if

they test positive or negative on the screener in a given setting. Posttest probabilities are calculated by multiplying a screener's likelihood ratio by the prevalence, also known as the base rate, of reading difficulties in a given setting; thus, posttest probabilities are considered sample-based statistics because they are dependent on the proportion of individuals with and without reading difficulties in a specific setting.

VanDerHeyden (2013) proposed that posttest probabilities of greater than or equal to .50 should be used to indicate need for intervention, while posttest probabilities of less than or equal to .10 should indicate that intervention or follow-up assessment be withheld. Posttest probabilities between .10 and .50 indicate insufficient confidence in a student's probability of reading risk and the need for follow-up assessment or intervention to improve this prediction. Additionally, for a screener to be deemed useful for instructional decision making, administering the screener should result in a meaningfully different probability of reading difficulties above and beyond the known base rate of reading difficulties in a setting.

Evidence of CBM Use for Universal Screening Purposes

Diagnostic accuracy research in education to date has generally focused on an examination of the diagnostic accuracy statistics related to the two primary screening purposes described above, aligning with diagnostic accuracy research conducted in psychology and medicine (e.g., Deeks & Altman, 2004; Moons & Harrell, 2003; Pepe, 2003; Swets, 1988). First, studies examine CBM use for accurately differentiating between students with and without a given condition (e.g., Smolkowski & Cummings, 2016), known in medicine as a screener's discriminative ability (Eusebi, 2013). Second, studies examine CBM use for predicting likelihood of reading difficulties in an individual

student (e.g., Van Norman et al., 2017; VanDerHeyden et al., 2018), described in medicine as the screener's predictive ability (Eusebi, 2013). An overview of this research is described below.

Evidence of Discriminative Ability. In medicine, a screener's discriminative ability is particularly important for making health policy decisions (Eusebi, 2013). In these cases, screeners are expected to provide an accurate estimation of how prevalent a condition is in a given population. A screener that is more accurate at differentiating between individuals with and without the condition will provide health officials with a quick indication of the types of interventions that need to be applied at a population level. In education, a screener's discriminative ability provides educators with an overall sense of how their reading system is functioning and alerts them to whether limited school resources should be dedicated toward shoring up core versus supplemental or intensive supports. A screener with high discriminative ability will also give educators confidence that their screener is accurately classifying most students, and that supplemental intervention supports are in general being funneled toward the students most in need.

Sensitivity and specificity values are key for evaluating a screener's discriminative ability, or the screener's ability to accurately differentiate between students with and without reading difficulties. As such, a screener's discriminative ability can be evaluated based on population-based statistics alone, which are not dependent on the base rate of reading difficulties in a given sample. Thus, sensitivity and specificity values should be generalizable across settings. Because of their generalizability, sensitivity and specificity rates are the most widely reported and studied diagnostic accuracy indices in evaluations of universal screeners. In a meta-analysis of diagnostic

accuracy studies evaluating oral reading fluency screeners, Kilgus et al. (2014) examined sensitivity and specificity rates across 34 oral reading fluency diagnostic accuracy studies for cut-scores where sensitivity rates were held at or above .80. Across these studies, a sensitivity rate of .80 or higher corresponded to a specificity rate of between .71 and .73. These findings suggest that across studies, oral reading fluency was able to accurately identify approximately 80% of students who actually had reading difficulties and 70% of students who did not have reading difficulties. Kilgus et al. (2014) interpreted these rates as evidence that oral reading fluency had reasonably accurate discriminative ability across studies. Though not reported, it can be inferred that had cut scores been set that prioritized a sensitivity rate of .90 or above as recommended by Jenkins et al. (2007), there would have been a resulting drop in specificity rates below acceptable values, as specificity rates drop with increases in sensitivity.

Studies examining the diagnostic accuracy of early literacy CBMs of nonsense word and real word reading are less prevalent (January & Klingbeil, 2020), and have indicated variable discriminative ability. In these studies, when holding sensitivity values at or above .90, specificity ranged from inadequate to acceptable (Catts et al., 2009; Clemens et al., 2011; Compton et al., 2010; Goffreda et al., 2009; January et al., 2016; Johnson et al., 2009; Smolkowski & Cummings, 2016). For example, in a comparison of early literacy screening measures, Clemens et al. (2011) reported low to acceptable specificity values, ranging from .52-.71, for measures of real word reading and letter naming fluency. In contrast, January et al. (2016) found widely acceptable specificity, ranging from .72- .88 for first grade students, and .73- .91 for second grade students for researcher-developed real and nonsense word reading measures.

Smolkowski & Cummings (2016) used different criteria to set cut scores, holding sensitivity at or above .80 on a measure of nonsense word fluency. In their study, specificity values ranged from .60- .81 across kindergarten through second grade. Thus, based on established sensitivity and specificity guidelines, most diagnostic accuracy evaluations have found that when cut scores are chosen for early literacy screeners that hold sensitivity values to an acceptable criteria for use in schools, their specificity values may be inadequate to borderline acceptable (Gersten et al., 2009; Jenkins et al., 2007; Johnson et al., 2009).

Evidence of Predictive Ability. While sensitivity and specificity rates are important for evaluating a screener's discriminative ability, they may not be sufficient for evaluating a screener's predictive ability (Eusebi, 2013). In these cases, screeners are expected to accurately estimate how likely an individual is to have a condition based on their screening result so that cost-benefit analyses can be made about the appropriateness of different treatment options as compared to doing nothing. In education, a screener's predictive ability provides a probability of individual students developing reading difficulties given a certain screening result. This information can help educators determine whether supplemental or intensive intervention is an appropriate next step for the student or whether it is preferable to administer follow up assessments or withhold provision of supplementary intervention. A screener with strong predictive ability will give educators confidence that they are making appropriate decisions about whether an individual student is in need of additional instructional supports.

Likelihood ratios are most appropriate for evaluating a screener's predictive ability because they provide a comparison between students who will fail and students

who will pass the end of year test. Thus, likelihood ratios provide a probability of student success or failure on an outcome measure given a specific screener result, an arguably more meaningful statistic in schools settings for predictive purposes (VanDerHeyden & Burns, 2018). Likelihood ratios have been less frequently reported within diagnostic accuracy research, but existing studies provide some evidence for the use of universal screeners for predictive purposes in school settings. In their meta-analysis of oral reading fluency studies, Kilgus et al. (2014) found positive likelihood ratios of 2.82 to 3.22 on average across studies, indicating a small increased likelihood of reading risk for positive test results, and negative likelihood ratios of .23 to .34, indicating a small decreased likelihood of reading risk for negative test results on average across studies.

In a more recent study, VanDerHeyden et al. (2018) examined the predictive accuracy of three commonly used screening measures in one suburban school district in the Midwest. Within a sample of 814 third grade students, the researchers examined how well each screening measure predicted likelihood of reading risk on a state accountability test at the end of the school year. In their study, an oral reading fluency screener failed to meet acceptable decision thresholds set by the National Response to Intervention Center for sensitivity (minimum acceptable value of .80). Additionally, the positive likelihood ratio for oral reading fluency in their study was 1.82, and the negative likelihood ratio was .48, indicating that neither positive nor negative test results on the oral reading fluency screener resulted in any meaningful change in the likelihood of reading difficulties for students who were administered the test.

Posttest probabilities are also an important indicator of a screener's predictive ability within school-based contexts because they provide a probability of a student's

likelihood of reading difficulties given a certain screening result that is specific to the prevalence of reading difficulties in that school context. In the VanDerHeyden et al. (2018) study, which had a sample base rate of reading difficulty of 16%, the corresponding positive posttest probability was 26% and negative posttest probability was 8% for oral reading fluency. In other words, in a context where 16% of students failed the end-of-year test, there was a 26% chance that students who failed the screening would also fail the end-of-year test, and there was an 8% chance that students who passed the screener would fail the end-of-year test. Thus, based on posttest probability recommendations (VanDerHeyden, 2013), the researchers determined that in a setting with a reasonably low prevalence rate of reading difficulties, oral reading fluency may do a poor job of helping to correctly classify students who fail a screening (e.g., are classified as at risk).

A keyword search of the ERIC electronic database using search terms *early literacy, CBM, curriculum-based measure, DIBELS, aimsweb, easyCBM, letter nam*, nonsense word*, word read*, likelihood ratio*, and *posttest** yielded no results, indicating that no studies have currently examined likelihood ratios and resulting posttest probabilities for early literacy CBM measures. However, given that likelihood ratios are calculated based on sensitivity and specificity values, and that the sensitivity and specificity values of early literacy screeners have been found to be on average poorer than oral reading fluency, early literacy CBMs are expected to likely also have poorer predictive ability than oral reading fluency.

In the context of Kane's argument-based approach to test validation, these studies provide evidence that early literacy CBM assessments range from inappropriate to

acceptable for discriminative purposes within MTSS-R. These studies also provide mixed evidence for oral reading fluency CBMs' acceptability for predictive purposes within MTSS-R, depending on the stakes of the instructional decisions being made (e.g., Kilgus et al, 2014; VanDerHeyden & Burns, 2018). However, these studies also point to the need for research examining early literacy CBMs' predictive ability. The variability in diagnostic accuracy statistics across these studies indicate the need for a systematic examination of specific contextual factors that may predictably alter early literacy CBMs interpretations and uses across different school contexts. A thorough examination of the extent to which correlational and diagnostic accuracy statistics vary depending on these contextual factors may result in a more nuanced understanding of how and when early literacy CBMs should be used for discriminative and predictive screening purposes within diverse school systems.

Lag Time as a Source of Bias in Diagnostic Accuracy Studies

Medical diagnostic accuracy research may elucidate how contextual factors might impact universal screener diagnostic accuracy in schools, as many parallels can be drawn between screening purposes in these two fields. For instance, in the field of medicine, screeners are used to classify individuals as having medical conditions in place of more intensive, expensive, and potentially invasive diagnostic tests (e.g. Steiner, 2003; Whiting et al., 2013). Similarly, in schools reading screeners are used in place of time intensive gold standard outcome measures to classify students as having reading difficulties. Additionally, screeners are used across settings for both discriminative and predictive purposes. Medical and educational screeners with strong discriminative ability are especially useful for making systems-level decisions such as how to allocate limited

public health or school-based resources, while screeners with strong predictive ability are most helpful for making decisions about individuals, such as what the appropriate course of treatment is for an individual with a chronic illness or with reading difficulties.

One important contrast, however, between medical and educational screeners is that medical screening assessments are generally intended to provide clinicians accurate information about whether or not an individual or group of individuals has a *current* medical problem. This enables the clinician to make informed decisions about whether or not to provide treatment or follow up testing. Diagnostic accuracy studies in the medical field are intentionally designed to assess screeners for this purpose. Accordingly, screening and outcome tests are administered in close temporal proximity to one another, a recommendation provided by two widely recognized assessments of the quality of diagnostic accuracy studies (QUADRAS-2; STARD Statement). These sources recommend minimizing the length of time between administrations of the screener and outcome measure within diagnostic accuracy studies because this "lag time" is recognized as one potential source of bias which may systematically alter diagnostic accuracy statistics from their "true" accuracy.

The addition of time between two test administrations may result in any number of unaccounted for contextual factors unique to the sample being studied contributing to systematic changes in individuals' condition over time. In these cases, individuals who were identified as having the condition by the screener may no longer have the condition when the outcome measure is administered, and vice versa. These unaccounted for changes may make a screener's diagnostic accuracy statistics less generalizable across settings (Bossuyt et al., 2015; Cohen et al., 2016; Whiting et al., 2004, 2011, 2013).

The issue of lag time bias presents a unique challenge when evaluating early literacy CBMs for use as screeners within MTSS-R, because the theoretical foundation of MTSS-R rests on the need to identify students at risk for developing reading difficulties in the *future* so that these students can be provided with early and effective intervention, thereby "ruining" these risk predictions (Baker et al., 2010; Gersten et al., 2009). In other words, in the context of MTSS-R, supplemental or intensive intervention is provided to students identified as at risk for reading difficulties at the start of the school year so that the intervention accelerates their rate of growth to such a degree that they no longer have reading difficulties at the end of the school year. Thus, MTSS-R systems rely on screeners to accurately predict *at the beginning of the year* whether or not students are expected to fail an *end-of-year* test of reading proficiency; in these contexts, screeners are expected to function as *prognostic* rather than *diagnostic* tests.

By extension, education researchers frequently study reading screener diagnostic accuracy in contexts where the screener is administered at the beginning of the school year and the outcome measure is administered at the end of the school year (e.g., Goffreda et al., 2009; Johnson et al., 2009; Petscher et al., 2011, Smolkowski & Cummings, 2016). In fact, education researchers have often been encouraged to include a gap between the two test administrations when evaluating universal screeners. For example, until their most recent call for academic screening tool evaluations, the National Center on Intensive Intervention (NCII) *required* that screening systems have a lag time of at least 3 months between administrations of the screener and outcome measure in order to be considered for evaluation (NCII, 2018). Further, though many diagnostic accuracy studies have examined diagnostic accuracy statistics across multiple time lags,

these statistics have often been discussed interchangeably without a discussion of any predictable differences in diagnostic accuracy statistics based on lag time.

To comprehensively evaluate early literacy CBMs for their two primary screening purposes, it is critical to determine whether lag time alters diagnostic accuracy statistics to such a degree that they cannot be expected to generalize across diverse school settings. One study in the field of education has explicitly evaluated the impact of lag time on diagnostic accuracy statistics. In their meta-analysis, Kilgus et al. (2014) considered how lag times of 0, 3, 6, 9, and over 12 months altered sensitivity, specificity, and likelihood ratio statistics across 34 diagnostic accuracy studies of oral reading fluency. They found great variation in lag time between studies, with most studies administering their outcome measure concurrently, within three months, or within six months of their screening measure. They also found that across the board studies demonstrated relatively stable sensitivity and specificity levels ranging from .74 to .83 for sensitivity and .71 and .77 for specificity across time lags. Positive and negative likelihood ratios were also fairly stable, with positive likelihood ratios ranging from 2.82 to 3.22 and negative likelihood ratios ranging from 0.23 and 0.34 across lag times. However, Kilgus et al. (2014) also found a slight systematic variation in diagnostic accuracy statistics based on lag time; sensitivity rates fell below what the authors identified as an acceptable level (< .80) when lag time was 9 months or more. Additionally, the cut score for risk associated with optimal sensitivity and specificity values varied greatly across studies. The Kilgus et al. (2014) did not report AUC values across studies, so an interpretation of whether overall accuracy varied across lag times is unavailable.

The Impact of Instructional Effectiveness on Lag Time

The Kilgus et al. (2014) meta-analysis provides initial evidence that lag time may impact diagnostic accuracy statistics to a small degree for an oral reading fluency screener and may consistently impact the cut-score associated with optimal sensitivity and specificity values. These findings demonstrate that diagnostic accuracy values may not necessarily be generalizable across school contexts. At the same time, the Kilgus et al. (2014) study did not examine the effect of lag time on the diagnostic accuracy of early literacy CBMs, suggesting the need for this research.

Further, the Kilgus et al. (2014) study failed to examine how key contextual factors within MTSS-R may differentially alter the impact of lag time on CBM diagnostic accuracy. One such critical factor is the effectiveness of instruction being provided to students within diagnostic accuracy studies. A primary goal of educators within MTSS-R is to systematically alter screening predictions for students who are identified as at risk for reading difficulties by providing these students with effective supplemental and intensive instruction. As such, it may be important to examine the extent to which the instruction being provided in the time between test administrations may impact a screener's diagnostic accuracy. Depending on how effective instruction is for students identified at risk, lag time may cause a screener to appear more or less accurate. For example, a screener would be expected to appear less accurate in a school setting where instruction successfully alters the risk category of many students who were classified as "at risk" on the beginning-of year screener than in a setting where instruction fails to alter these students' risk category. The effect would be expected to grow more pronounced as lag time increased. To effectively determine the accuracy of an early literacy screener

within MTSS-R then, instructional effectiveness is a key contextual factor that must be explicitly examined.

Researchers in the medical world have begun to consider the extent to which treatments may alter diagnostic accuracy statistics. In a systematic review of diagnostic accuracy evaluations in medicine, Whiting et al. (2013) describe this "treatment paradox" as one potential source of diagnostic accuracy bias. The treatment paradox is defined as any instance in which a treatment is initiated for an individual based on screening results and the outcome test is administered following the treatment (Whiting et al., 2004). In their review, Whiting et al. (2013) found only one meta-review which studied the treatment paradox; this meta-review found no systematic differences in diagnostic accuracy estimates based on whether treatment was provided. However, many studies within the review provided no treatment or failed to report whether treatment was provided (Rutjes et al., 2006), suggesting the need for more research to examine the potential impact of the treatment paradox on diagnostic accuracy statistics.

Though an ERIC keyword search using search terms *diagnostic accuracy, sensitivity, specificity,* and *instruction** produced no studies examining how the treatment paradox, in the guise of instruction, may alter the diagnostic accuracy of screeners in education, findings from existing literacy intervention studies suggest a need for this research. For example, numerous randomized controlled trials demonstrate that a majority of at risk students who are provided with systematic and explicit supplemental reading interventions accelerate their rate of learning such that they fully or nearly catch up to their classmates who were not at risk and did not receive the intervention. These students also improve their skills above and beyond at risk students who have not

received the supplemental intervention (Gersten, Newman-Gonchar, et al., 2017; Wanzek et al., 2016).

These findings indicate that when provided with effective supplemental intervention, many students who would have otherwise failed an end-of-year reading test will instead pass the end-of-year test because their reading skills have improved; this effect would be expected to grow stronger with increased time spent in supplemental intervention. In these cases, lag time may alter diagnostic accuracy statistics most in contexts with highly effective supplemental instruction, where a large proportion of at-risk students would change reading status.

Base Rate of Reading Difficulties. Posttest probabilities may be especially impacted by differences in supplemental instructional effectiveness because they are highly dependent on the base rate of reading difficulties in a setting. It is widely acknowledged that screeners with similar sensitivity and specificity values will produce different posttest probability values based on the sample base rate of reading difficulties (e.g., Petscher et al., 2011; Van Norman et al., 2017; VanDerHeyden et al., 2018). Regardless of sensitivity and specificity values, as base rate increases in a sample, posttest probabilities for positive screening results increase and posttest probabilities for negative screening results decrease.

Yet researchers have largely regarded a school's base rate as static across the year, which is problematic given that effective MTSS-R systems have been shown to successfully alter the proportion of students at risk for reading difficulties across a school year. The impact of shifting base rates may be especially important to consider in early elementary school when students' skills are expected to rapidly develop and change

(Speece, 2005). In order to make accurate recommendations about the extent to which early literacy CBMs are appropriate for both discriminative and predictive purposes in early elementary school, it is critical to explicitly study the extent to which supplemental instructional effectiveness differentially alters the base rate of reading difficulties, and thus posttest probabilities, in these early grades.

Statement of the Problem

Reading screeners are most often used within the context of MTSS-R to (1) evaluate the current effectiveness of a school's reading system and (2) assign students to supplemental instruction to prevent future reading difficulties. For evaluating reading systems, educators rely on screeners to accurately discriminate between students with and without current reading difficulties and make decisions about whether to dedicate limited school resources to core versus supplemental or intensive instruction. For assigning students to supplemental instruction, educators use screeners to accurately predict students' likelihood of having future reading difficulties. They then assign students who are at risk for reading difficulties to supplemental instruction intended to move these students from at risk to proficient readers, thereby ruining their screener predictions so that "every struggling reader becomes a false positive" (K. Smolkowski, personal communication, April 23, 2020).

Using an argument-based approach to test validation, a comprehensive screener evaluation within MTSS-R should prioritize examining specific diagnostic accuracy statistics that align with each of these two purposes. For evaluating current systems effectiveness, it is essential to consider a screener's ability to differentiate between proportions of students who do and do not have *current* reading difficulties. For this

purpose, then, it is most important to examine sensitivity and specificity values for a screener predicting to an outcome measure administered at the same time point. In contrast, for assigning students to supplemental intervention it is necessary to consider a screener's ability to accurately predict an individual student's likelihood of having reading difficulties *at the end of the school year*. In this case, it is most imperative to examine a screener's likelihood ratios and posttest probabilities when predicting to an outcome measure administered at the end of the school year.

In the field of education to date, diagnostic accuracy studies evaluating early literacy CBMs have generally focused on examining sensitivity and specificity values alone (e.g., Clemens et al., 2011; Johnson et al., 2009; Smolkowski & Cummings, 2016). Further, these studies have reported these sensitivity and specificity values across lag times between screeners and outcome measures interchangeably, without consideration of whether diagnostic accuracy statistics vary based on the amount of time between the two test administrations. Further research is needed to determine the extent to which lag time matters when evaluating early literacy CBMs as universal screeners in the context of MTSS-R, such that findings from existing screening evaluations can be interpreted for educator use in a more nuanced manner.

Additionally, an argument-based approach to screener test validation calls for the need to consider instructional context when evaluating screeners for their intended purposes within MTSS-R. Many education researchers who have conducted these diagnostic accuracy evaluations have conjectured that instruction may impact diagnostic accuracy statistics in some way (e.g., Smolkowski & Cummings, 2016; Petscher et al., 2011; VanDerHeyden, 2013). However, the potential impact of this "treatment paradox"

has not yet been explicitly examined in the field of education and warrants further study. Instructional effectiveness may impact both the discriminative and predictive ability of a reading screener. In particular, the time interval between the screener and outcome measure may differentially alter diagnostic accuracy statistics within a study based on the effectiveness of the supplemental instruction being provided to at risk students.

In the field of education, most diagnostic accuracy studies have typically been conducted within diverse instructional contexts, and the quality and content of instruction received by students who are classified as at risk and not at risk in these studies is typically not reported. Widely varying instructional contexts and study populations make it difficult for educators to evaluate the relative benefits of screening tools in their own instructional contexts for both discriminating between students who currently do and do not have reading difficulties, as well as for predicting the likelihood of an individual student demonstrating future reading difficulties. Without reporting on the instruction that is being provided within a diagnostic accuracy study, it may be difficult to generalize study findings for use in contexts with varying instructional effectiveness, as is typically seen in schools implementing MTSS-R.

The Current Study

Studying diagnostic accuracy statistics within the context of a randomized controlled trial may help to illustrate how an early literacy CBM may have varying test score interpretations and uses in different instructional settings with increased lag time between administrations of screener and outcome measures. Within this context, a clearer comparison of the accuracy of the screening tool can be made between settings in which all variables are controlled for apart from the instruction being provided, allowing

educators to better estimate how useful the tool will be in their own setting for discriminative and predictive purposes.

Using the context of a randomized controlled trial, the current study aims to explore how lag time may result in the treatment paradox, and thus alter a reading screener's test score interpretations and uses for discriminative and predictive purposes within MTSS-R settings, based on the relative effectiveness of instruction being provided to students. Though experts in the medical field have warned that the introduction of an effective treatment may systematically alter diagnostic accuracy values (e.g., Cohen et al., 2016), this is the first study in the field of education that uses the context of a randomized controlled trial to examine the impact of effective instruction on a reading screener's concurrent and predictive correlations, overall accuracy, sensitivity, specificity, likelihood ratios, and posttest probability values.

The present study examines how lag time and instructional effectiveness may impact the evaluation of the DIBELS 6th Edition Nonsense Word Fluency measure for two primary uses: (1) for discriminating between students who do and do not have reading difficulties for evaluating the current effectiveness of a school's reading system and (2) for predicting an individual student's likelihood of future reading difficulties. The current study conducts a series of correlational and diagnostic accuracy analyses to examine the discriminative and predictive ability of Nonsense Word Fluency predicting to two different outcome measures. These analyses are conducted in the context of schools which were randomly assigned to receive either a highly explicit and systematic reading intervention within the context of MTSS or to a business-as-usual MTSS comparison group.

Instructional effectiveness was defined in the current study based on school assignment to study condition in the original ECRI study. Original study condition was deemed an appropriate proxy for instructional effectiveness in the current study based on findings from the original ECRI study which demonstrated strong treatment effects for at risk (Tier 2) students in the ECRI treatment group on measures of nonsense word fluency, oral reading fluency, and untimed real and nonword reading. Specifically, in the original ECRI study, from fall to winter of first grade, Tier 2 students in the treatment condition outperformed Tier 2 comparison students to a statistically significant degree on measures of letter-sound correspondence (NWF-CLS; $g = 0.31$) and word blending skills (NWF-WRC; $g = 0.37$), and to a marginally significant degree on a measure of oral reading fluency (ORF; $g = 0.20$). From fall to spring of first grade, Tier 2 students in the treatment condition outperformed Tier 2 comparison students to a statistically significant degree on timed measures of letter-sound correspondence (NWF-CLS; $g = 0.39$), word blending (NWF-WRC; $g = 0.41$), and oral reading fluency (ORF; $g = 0.25$), and on untimed measures of real word (WRMT Word ID; $g = 0.41$) and nonword (WRMT Word Attack; $g = 0.48$) reading. Tier 2 students in the treatment condition also outperformed Tier 2 students in the comparison condition on a standardized test of total reading (SAT-10 Total Reading; $g = 0.12$), word reading (SAT-10 Word Reading; $g = 0.06$), and sentence reading (SAT-10 Sentence Reading; $g = 0.01$), though these values were not statistically significantly different (Fien et al., 2020). Thus, it can be argued that because instruction was more effective overall for Tier 2 students in the original ECRI study, all students in the ECRI treatment condition in the current study were in an instructional context with “higher instructional effectiveness”, while all students in the comparison

condition in the current study were in an instructional context with "lower instructional effectiveness".

Within this context, an argument-based approach to test validation is used to compare Nonsense Word Fluency's test score interpretations and uses for predictive and discriminative purposes across lag times and between schools providing more and less effective instruction (e.g. ECRI intervention vs. business-as-usual comparison condition) to illustrate how provision of evidence-based intervention between administrations of a screener and two different outcome measures may alter Nonsense Word Fluency's diagnostic accuracy for these purposes.

Research Questions

This study aims to address four primary research questions:

Research Question 1: What is the evidence for an early literacy CBM's test score interpretations within the context of MTSS-R? (i.e., What are the concurrent and predictive correlations for scores associated with a measure of decoding skills (NWF-CLS) relative to a test of oral reading fluency (ORF) and a multiple-choice test of overall reading achievement (SAT-10) in first grade?

- Research Question 1a: Does the evidence for an early literacy CBM's test score interpretations vary based on (1) time of year and (2) lag time between test score administrations?

Research Question 2: What is the evidence for an early literacy CBM's discriminative ability within the context of MTSS-R? (i.e., What are the Area Under the Curve (AUC), sensitivity, and specificity values for a measure of decoding skills (NWF-CLS) predicting

proportions of students with and without reading difficulties on a test of oral reading fluency (ORF) and a multiple-choice test of overall reading achievement (SAT-10)?)

- Research Question 2a: Does the evidence for an early literacy CBM's discriminative ability meaningfully differ based on (1) time of year and (2) lag time between test score administrations?

Research Question 3: What is the evidence for an early literacy CBM's predictive ability within the context of MTSS-R? (i.e., What are the positive and negative likelihood ratios and positive and negative posttest probabilities for a measure of decoding skills (NWF-CLS) predicting individual students' likelihood of reading difficulties on a test of oral reading fluency (ORF) and a multiple-choice test of overall reading achievement (SAT-10)?

- Research Question 3a: Does the evidence for an early literacy CBM's predictive ability meaningfully differ based on (1) time of year and (2) lag time between test score administrations?

Research Question 4: Does the evidence for an early literacy CBM's test score interpretations and discriminative and predictive uses within the context of MTSS-R meaningfully differ based on a setting's instructional effectiveness? (i.e., do concurrent/predictive correlations, AUCs, sensitivity and specificity values, positive and negative likelihood ratios, and positive and negative posttest probabilities meaningfully differ between the ECRI treatment condition and the business-as-usual comparison condition?)

II. METHOD

This study analyzed data from a large-scale cluster randomized controlled trial aimed at evaluating the efficacy of first grade Enhanced Core Reading Instruction (ECRI), a multitiered reading intervention (Fien et al., 2015; Smith et al., 2016). ECRI was developed to improve on educators' use of explicit and systematic instructional principles during reading instruction. In the original ECRI study, 44 schools in 9 districts in Oregon and Massachusetts were recruited across two waves and participated in the study for two years, for a total of 8,808 1st grade students and their teachers who were nested within schools. Schools were eligible to participate in the larger ECRI study if they (a) used a published core reading program during a 90-minute Tier 1 reading block and (b) provided Tier 2 students with 30 minutes of daily small-group instruction.

Participants

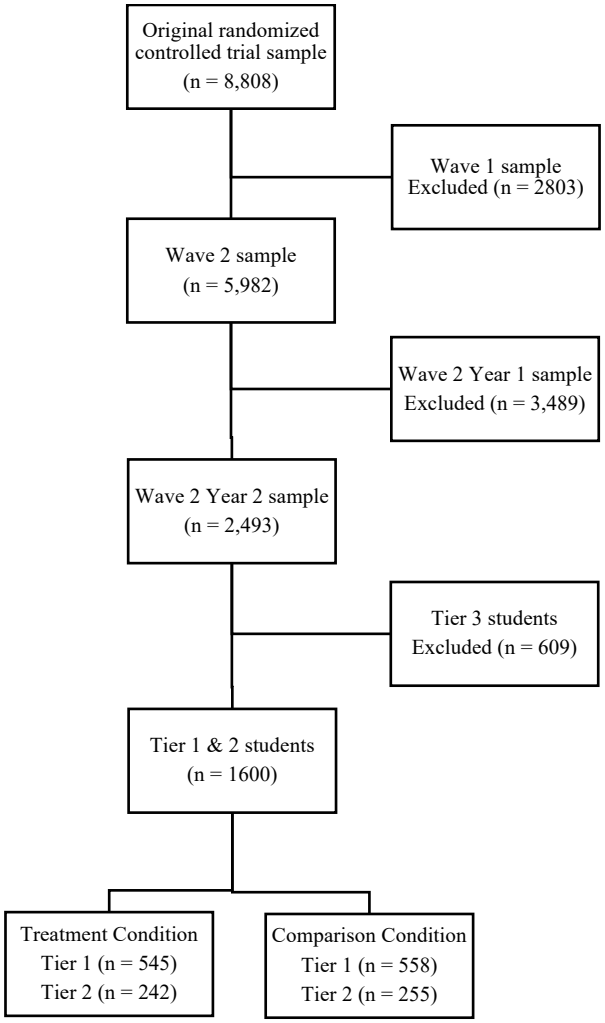
The current study analyzed Tier 1 and Tier 2 student data from the second year of ECRI implementation for Wave 2 schools, which were comprised of 20 schools in three districts in Oregon and eight schools from three districts in Massachusetts. In the original study, fall SAT-10 percentile ranks based on normative criteria from the SAT-10 (2007) technical manual were used to assign students to tiers of instruction. Students who scored between the 10th and 30th percentile were assigned to receive both Tier 1 instruction and Tier 2 intervention, while students above the 30th percentile were assigned to receive Tier 1 instruction alone. Due to the original study design, data for all students in Wave 1, for all students in Year 1 of Wave 2, and for students assigned to Tier 3 instruction in Year 2 of Wave 2 were unavailable for the current analyses based on the data collected in the original ECRI study. In other words, all Wave 1 students, Wave 2 students who

participated in the first year of the randomized controlled trial, and Wave 2 students who participated in the second year of the randomized controlled trial but who performed below the 10th percentile on fall SAT-10 were not included in analyses. Participants in the current study included 1600 first grade students assigned to Tier 1 or Tier 2 who attended either a treatment (N = 787) or comparison (N = 813) school in the ECRI study. Figure 1 illustrates the final sample for the current study, including exclusion criteria based on wave, study, or tier of instruction.

A total of 1103 students were assigned to Tier 1 instruction and included in study analyses (545 in treatment; 558 in comparison). An additional 497 students were assigned to Tier 2 intervention and included in study analyses (242 in treatment; 255 in comparison). For students included in the current study, 2.9% received special education services (2.5% in treatment; 3.0% in comparison) and 11.4% were English Learners (15.5% in treatment; 7.4% in comparison). Though race/ethnicity data were unavailable for students in the current study, data from the National Center for Educational Statistics (NCES, 2011) indicated that for schools participating in the original ECRI study, approximately 19.8% of students identified as Hispanic (22.8% in treatment; 16.9% in comparison), and 3.9% of students identified as African American (4.8% in treatment; 3.0% in comparison). Approximately half of students (50.3%) were eligible for free or reduced-price lunch (54.5% in treatment; 46.0% in comparison). A total of 99 teachers participated in the current study. Teachers reported an average of 14.30 years of teaching experience (SD = 10.03); total years of teaching experience was similar between treatment (M = 13.52, SD = 9.57 years) and comparison (M = 15.14, SD = 10.45 years) conditions.

Figure 1

Flow Chart Illustrating Current Study Sample



Instruction Implementation

In both treatment and comparison conditions, teachers provided daily reading instruction using a comprehensive core reading program to all students during a 90-minute core reading block. Students identified as needing Tier 2 supports were administered an additional 30 minutes of daily small group reading instruction.

Treatment Condition

The ECRI multitiered intervention was designed to increase (1) the quality and explicitness of instruction provided in Tiers 1 and 2 through the use of lesson maps that prioritized critical content from the core reading program, (2) the specificity of instructional materials in Tiers 1 and 2 through the use of explicit teaching routines, and (3) the alignment between Tier 1 and 2 instruction. The ECRI intervention includes Tier 1 enhanced core reading instruction, Tier 2 small group instruction, and initial and ongoing professional development and coaching. Additionally, the intervention includes the use of data-based decision making to inform instructional changes within and across tiers of instruction throughout the school year. Students are initially placed in Tier 2 small group instruction based on initial skill and regrouped as needed throughout the year based on data. Thus, the ECRI intervention emphasizes and reinforces key components of high-quality MTSS-R, including screening and progress monitoring, evidence-based instruction and intervention, ongoing data-based decision making, and ongoing professional development and coaching aimed at increasing teachers' fidelity of implementation. Fidelity of implementation observations in treatment classrooms indicated that the mean score for the quality of explicit instruction was 0.89 (*SD*= 0.17).

Comparison Condition

Teachers in the comparison condition provided core instruction through the use of an adopted core reading program. These teachers reported that during the core instructional block, they spent an average of 52.5 (*SD* = 31.0) minutes in whole group instruction, 34.5 (*SD* = 26.3) minutes in small group instruction and 27.9 (*SD* = 15.6) minutes in independent work. Tier 2 instruction in comparison schools varied, with

teachers reporting that Tier 2 instruction included a variety of published, standardized protocol intervention materials and teacher-developed materials. 62% of teachers in the comparison condition also indicated that they received some degree of literacy-related professional development and coaching. Fidelity of implementation observations indicated that the mean quality of explicit instruction in comparison classrooms was 0.49 (*SD* = 0.25). Thus, data suggest that on average, teachers in ECRI classrooms provided higher quality explicit instruction targeting foundational early literacy skills.

Measures

DIBELS 6th Edition Nonsense Word Fluency (NWF)

DIBELS Nonsense Word Fluency (University of Oregon, 2002) is an individually-administered, timed fluency measure of students' decoding skills. Students are asked to read from a list of consonant-vowel and consonant-vowel-consonant pseudowords for one minute. Students can either read the words sound-by-sound or as whole words. Performance on Nonsense Word Fluency results in two scores: Correct Letter Sounds, which provides a measure of letter-sound correspondence and is calculated by counting the number of correct individual letter sounds the student produces, and Words Recoded Correctly, which provides a measure of word blending skills and is calculated by counting the number of words the student reads correctly as a whole word. Alternate form reliability for Nonsense Word Fluency ranges from .67 to .80. Concurrent validity coefficients range from .35 to .55 when comparing Nonsense Word Fluency to the readiness subtests of the Woodcock-Johnson Psycho-Educational Test (University of Oregon, 2002). Preliminary analyses indicated similar patterns of results for both Nonsense Word Fluency- Correct Letter Sounds and Nonsense Word

Fluency- Words Recoded Correctly in the current study, with Nonsense Word Fluency- Correct Letter Sounds demonstrating slightly stronger diagnostic accuracy across outcome measures and lag times. Therefore, for the purpose of the current study results are reported for Nonsense Word Fluency- Correct Letter Sounds scores only and Nonsense Word Fluency- Correct Letter Sounds is referred to as Nonsense Word Fluency.

DIBELS 6th Edition Oral Reading Fluency (ORF)

DIBELS Oral Reading Fluency (University of Oregon, 2002) is an individually-administered, timed fluency measure of students' skill with reading connected text accurately and fluently. Students are presented with three short passages and asked to read each passage aloud for one minute. The final score produced for each passage is the number of words read correctly in one minute. A student's benchmark score is determined by taking the median score from the three passages. Oral Reading Fluency demonstrates strong alternate-form and test-retest reliability, with coefficients ranging from .89 to .94, and .92 to .97, respectively (University of Oregon, 2002). DIBELS Oral Reading Fluency has also demonstrated strong predictive validity with reading comprehension measures, with coefficients ranging from .65 to .80 (Roehrig et al., 2008; Shapiro et al., 2008).

Stanford Achievement Test Series, 10th Edition (SAT-10)

The SAT-10 (Harcourt Educational Measurement, 2002) is a group-administered, standardized test of reading achievement. The SAT-10 is untimed, and students are asked to answer a series of multiple-choice questions to assess their skills in a variety of foundational skills, including recognizing sounds and letters, word reading, and reading

comprehension. For the purpose of the current study, trained data collectors administered the appropriate versions of the SAT-10 in the fall and spring of 1st grade to all students. In the fall, they administered the Stanford Early School Achievement Test (SESAT) 2, which is comprised of the Sounds and Letters, Word Reading, and Sentence Reading subtests. In the spring, they administered the Primary 1, which is comprised of the Word Study Skills, Word Reading, Sentence Reading, and Reading Comprehension subtests. Testing time ranged from 110 to 155 minutes across administrations. According to the test manual, the internal consistency reliability coefficient is .94 for the SESAT 2 and .97 for the Primary 1. Total Reading scores for both the SESAT 2 and Primary 1 are correlated with the Otis-Lennon School Ability Test, 8th Edition, Total scores (r = .68 and .61, respectively). In the present study, the percentile rank associated with the scale score for the total reading domain was used for analysis.

Procedures

DIBELS 6th Edition and SAT-10 data were collected by trained data collectors. DIBELS 6th Edition Nonsense Word Fluency was administered to all participating students in the fall, winter, and spring of 1st grade. DIBELS 6th Edition Oral Reading Fluency was administered to all students in the winter and spring of 1st grade, and SAT-10 was administered to all students in the fall and spring of 1st grade. Data collectors participated in three initial days of data collection training in the fall prior to the start of data collection and four additional days of training across the winter and spring. Inter-rater reliability data was collected for individually administered measures by assessment coordinators who shadow scored assessors. Average inter-rater reliability was 92.9%

(range = 87-110%) for Nonsense Word Fluency and 97.9% (range = 94-100%) for Oral Reading Fluency across the study.

Analyses

IBM SPSS Statistics 26 was used for data processing and analysis. Preliminary analyses were conducted to inspect the data for out of range values and missing data. A preliminary analysis of distributional properties of Nonsense Word Fluency (NWF-CLS), Oral Reading Fluency (ORF), and SAT-10 scores at each timepoint was conducted to see if they met assumptions and were normally distributed.

Research Question 1 and 1a: Overall Test Score Interpretations

To answer Research Question 1 and 1a, Pearson's *r* bivariate correlations were calculated for the overall sample for each combination of screening (i.e., Nonsense Word Fluency) and outcome (i.e., Oral Reading Fluency, SAT-10) measure at each time point available (i.e., fall, winter, or spring).

Research Question 1 and 1a Hypotheses. Regarding Research Question 1, it was hypothesized that across time points, Nonsense Word Fluency would be highly correlated with Oral Reading Fluency and moderately correlated with the SAT-10. Regarding Research Question 1a, it was predicted that for the overall sample, concurrent correlations would be similar in the fall, winter, and spring, and that predictive correlations would be smaller than concurrent correlations across measures, with correlations decreasing with increased lag time between test administrations.

Research Question 2 and 2a: Overall Discriminative Ability

To answer Research Question 2 and 2a, classification accuracy was examined for the overall sample via a series of Receiver Operator Curve (ROC) analyses using

methods described by Smolkowski et al. (2015). ROC curves were generated examining the diagnostic accuracy of Nonsense Word Fluency for predicting to the 40th normative percentile on the SAT-10 in the fall and spring and on Oral Reading Fluency in the winter and spring.

The 40th percentile was used as the cut-off for identifying students who have reading difficulties for two reasons. First, performance at or above the 40th percentile is frequently used to classify students who are at low risk on state achievement tests (American Institutes for Research, 2007) and for federal reporting (e.g., Reading First). Any students who fall below the 40th percentile can be deemed at some risk for reading difficulties and in need of supplemental reading supports. Finally, because data were not available for students assigned to Tier 3 intervention in the original ECRI study, choosing a cut-off that corresponded to a lower percentile rank, such as the 20th or 30th percentile, in the current study would have resulted in a proportion of students that was too small to provide precise diagnostic accuracy estimates for the Nonsense Word Fluency screener predicting to each outcome measure.

From these ROC curves, the Area Under the Curve (AUC) was examined as an overall indicator of classification accuracy. An optimal cut score for risk on Nonsense Word Fluency was then identified for each combination of screener and outcome measure at each time point and across lag times using a commonly applied decision rule in education research which prioritizes identifying a majority of students with failing scores on the outcome measures. Specifically, a cut score for risk was selected by identifying the cut score that corresponds to the highest possible specificity value when sensitivity was .90 or higher, an approach recommended by Jenkins et al. (2007) based on the argument

that it is more justifiable to provide supplemental intervention supports to students who may not need them than to fail to provide supports to students in need.

For all analyses, AUC, sensitivity and specificity values included 95% confidence bounds. Confidence bounds were calculated using the score method corrected for continuity (Newcombe, 1998). AUCs, sensitivity, and specificity values were then examined across each time point, lag time and outcome measure for non-overlapping confidence bounds. Values with non-overlapping confidence bounds were interpreted at meaningfully different.

Research Question 2 and 2a Hypotheses. Regarding Research Question 2, it was hypothesized that across time points and lag times, diagnostic accuracy statistics associated with Nonsense Word Fluency's discriminative ability (i.e., AUC, sensitivity and specificity) would be inadequate based on the Jenkins et al. (2007) recommendation to hold sensitivity at or above .90. Specifically, it was hypothesized that when sensitivity was held at .90 or higher, specificity values would fall between .40-.60, as indicated by prior research (Clemens et al., 2011; Johnson et al., 2009).

Regarding Research Question 2a, it was hypothesized that for both Oral Reading Fluency and SAT-10, Nonsense Word Fluency's discriminative ability would meaningfully differ across time lags (i.e., fall to spring, winter to spring, spring to spring), contrary to findings by Kilgus et al. (2014). This is because in the current sample, many students identified as at risk on fall SAT-10 were provided with daily Tier 2 instruction, and so it was expected that many of these students would transition to the population of students without reading difficulties as the year progressed. Specifically, it

was expected that as lag time increased, and when holding sensitivity values at .90 or higher, AUC and specificity values would decrease.

Research Question 3 and 3a: Overall Predictive Ability

To answer Research Question 3 and 3a, positive and negative likelihood ratios and posttest probabilities for each ROC curve analysis were calculated using a Diagnostic Test Calculator (Schwartz, 2021), and 95% confidence bounds were calculated for positive and negative likelihood ratios using the score method corrected for continuity (Newcombe, 1998). Posttest probabilities were calculated for each ROC curve analysis using the base rate of reading difficulties (ρ) specified by the target outcome measure as the pretest probability value (i.e., pre-test risk for reading difficulties). Positive and negative likelihood ratios were also examined across each time point, lag time and outcome measure for non-overlapping confidence bounds. Non-overlapping confidence bounds were interpreted at meaningfully different.

Research Question 3 and 3a Hypotheses. Regarding Research Question 3, it was predicted that diagnostic accuracy statistics associated with Nonsense Word Fluency's predictive ability (i.e., likelihood ratios, posttest probabilities) would show small increases or decreases in likelihood of having reading difficulties (e.g. 1.50 to 3.00 for positive likelihood ratios and .20 - .50 for negative likelihood ratios; posttest probabilities slightly greater than .10 for negative test results and slightly less than .50 for positive test results), and would generally not be acceptable for decision-making in schools given similar prior research on oral reading fluency CBMs (e.g. Kilgus et al., 2014; VanDerHeyden et al., 2018; Van Norman et al., 2017).

Regarding Research Question 3a, it was hypothesized that sample-based statistics (i.e., posttest probabilities) but not population-based statistics (i.e., likelihood ratios) would meaningfully differ across concurrent administration timepoints (i.e., fall to fall, winter to winter, spring to spring). Specifically, it was predicted that as the year progressed, positive posttest probabilities would become increasingly weaker and less useful for decision making, while negative posttest probabilities would grow increasingly stronger and more useful for decision making. This is because it was expected that as the number of students with reading difficulties decreased across the school year due to the provision of daily Tier 2 intervention, so would the overall base rate of reading difficulties. Likelihood ratios were not expected to meaningfully differ for concurrent analyses across timepoints or outcome measures because they are thought to be minimally impacted by base rate (e.g., Smolkowski & Cummings, 2015). Finally, both likelihood ratios and posttest probabilities were expected to meaningfully decrease with increased lag time (i.e., spring to spring vs winter to spring vs fall to spring) due to expected decreases in AUC and specificity values as lag time increased.

Research Question 4: Validity by Instructional Effectiveness

To answer Research Question 4, concurrent and predictive correlations and discriminative and predictive diagnostic accuracy values across time points, lag times, and outcome measures were calculated and compared for the treatment (i.e., higher instructional effectiveness) and comparison (i.e., lower instructional effectiveness) conditions, and examined for meaningful differences. ROC curves were visually compared, and AUCs, sensitivity and specificity values, and likelihood ratios were

examined for overlap in confidence bounds. Estimates that did not overlap with the confidence bounds around other estimates were considered meaningfully different.

Research Question 4 Hypotheses.

Test Score Interpretations. It was hypothesized that concurrent correlations would be similar across treatment and comparison conditions, while fall and winter screening correlation coefficients predicting to spring outcome measures would be smaller in the treatment (i.e., higher instructional effectiveness) condition than in the comparison (i.e., lower instructional effectiveness) condition.

Test Score Uses: Discriminative Ability. It was hypothesized that diagnostic accuracy statistics associated with Nonsense Word Fluency's discriminative ability (i.e., AUCs, sensitivity and specificity values) would appear less accurate and thus less useful for decision making in the higher instructional effectiveness condition than in the lower instructional effectiveness condition as the lag time between screener and outcome measures increased, while concurrent diagnostic accuracy values (i.e., fall to fall, winter to winter, spring to spring) would be comparable to one another.

Test Score Uses: Predictive Ability. It was hypothesized that for concurrent test administrations, posttest probabilities would grow progressively less useful for ruling in reading difficulties and more useful for ruling out reading difficulties in the higher instructional effectiveness condition than the lower instructional effectiveness condition as the year progressed. This was because it was expected that there would be a greater decrease in base rate of reading difficulties across the year in the treatment condition due to higher instructional effectiveness.

III. RESULTS

Table 1 shows descriptive statistics for all screening and outcome measures at each time point. For the overall sample, Tier 1 and 2 students grew an average of 31.50 points from fall to winter, and 46.97 points from fall to spring on Nonsense Word Fluency. Students in the overall sample grew an average of 23.78 points from winter to spring on Oral Reading Fluency and dropped an average of 2.82 percentage points from fall to spring on SAT-10 Total Reading percentile rank.

In the ECRI condition, Tier 1 and 2 students grew an average of 33.69 points from fall to winter, and 49.53 points from fall to spring on Nonsense Word Fluency. These students grew an average of 21.21 points on Oral Reading Fluency from winter to spring. The average SAT-10 percentile rank for all Tier 1 and 2 students in the ECRI condition dropped 3.92 percentage points from fall to spring, with an average percentile rank of 59.00 (*SD* = 26.35) in the fall and 55.08 (*SD* = 22.98) in the spring.

In the comparison condition, Tier 1 and 2 students grew an average of 29.57 points from fall to winter, and 44.00 points from fall to spring on Nonsense Word Fluency. These students grew an average of 23.24 points on Oral Reading Fluency from winter to spring. The average SAT-10 percentile rank for students in the comparison condition dropped 2.25 percentage points from fall to spring, with an average percentile rank of 57.69 (*SD* = 25.97) in the fall and 55.44 (*SD* = 23.75) in the spring.

Missing Data

Potential differences in reading scores for students with missing data were tested using missing at random procedure. Fall data from all students in the current study sample were analyzed using a Welch's *t'*-test for independent observations to determine

Table 1

Descriptive Statistics For Screener and Outcome Measures

		Fall			Winter			Spring		
		Overall	Treatment	Comparison	Overall	Treatment	Comparison	Overall	Treatment	Comparison
NWF-CLS	M	56.00	57.19	55.12	87.50	90.88	84.69	102.97	106.72	99.12
	(SD)	(30.77)	(30.81)	(30.09)	(37.42)	(34.91)	(39.24)	(32.08)	(30.30)	(33.47)
	N	1499	726	722	1462	707	754	1499	736	763
ORF	M	-	-	-	63.59	65.89	61.64	87.37	87.10	84.88
	(SD)		-	-	(38.37)	(36.80)	(39.48)	(35.95)	(33.94)	(37.04)
	N				1459	706	752	1495	735	760
SAT-10	M	58.06	59.00	57.69	-	-	-	55.24	55.08	55.44
	(SD)	(26.14)	(26.35)	(25.97)	-	-	-	(23.37)	(22.98)	(23.75)
	N	1601	787	813				1523	744	778

Note. ECRI = Enhanced Core Reading Instruction; NWF-CLS = Nonsense Word Fluency- Correct Letter Sounds; ORF = Oral Reading Fluency; SAT-10 = Stanford Achievement Test, 10th Edition Total Reading Percentile Rank.

whether data were missing at random. There were no significant differences on any of the fall reading measures for students with and without missing data; Nonsense Word Fluency *t'*(1497) = 0.62, *p* = .54, 95% CI [-4.35, 8.33], SAT-10 *t'*(1599), *p* = .07, 95% CI [-13.95, 0.41]. Similarly, fall data for participating students were analyzed using a Welch's *t'*-test for independent observations to determine whether there were significant differences in reading scores between students in the treatment and comparison conditions, and revealed that there were no statistically significant differences between conditions; Nonsense Word Fluency *t'*(1496) = -1.38, *p* = .17, 955 CI [-5.32, 0.92], SAT-10 *t'*(1598) = -1.19, *p* = .24.

Research Question 1 and 1a: Overall Test Score Interpretations

To answer Research Question 1, the relations between each screening and outcome measure were examined with Pearson's correlation analyses for the overall sample. Correlation analyses assume linearity, normality, homoscedasticity, and independence of errors. Scatterplots, histograms, and P-P plots were examined for the relation between each pair of screening and outcome measures at each time point. Scatterplots, histograms, and P-P plots are displayed in Figures 14-16 in the Appendix.

For all combinations of measures and timepoints, scatterplots presented a moderate, positive, linear relationship, suggesting that the assumption of linearity was tenable. The scatterplot of standardized predicted values also demonstrated that the data met the assumption of linearity. The scatterplot of standardized predicted values suggested that the data met the assumption of homoscedasticity for some but not all pairs of measures. For both Oral Reading Fluency and SAT-10 at fall, winter, and spring timepoints, errors frequently appeared to be heteroscedastic, with variance of residuals

decreasing as x-axes increased in many cases. This suggests that the assumption of homoscedasticity across all measures may not be tenable. Histograms of the standardized residuals for each combination of measures and timepoints indicated that the data had approximately normally distributed errors; P-P plots provided further evidence of normally distributed errors, demonstrated by points that were close to though not always entirely on the line.

Correlations were examined for all measures at all timepoints for the overall sample and are displayed in Table 2. The size of each correlation was interpreted based on Cohen (1988)'s benchmarks, suggesting that $|r| = .1, .3$, and. 5 indicate a small, medium, and large correlation, respectively. All correlations were significant ($p < .01$).

Correlations between each combination of measures were moderate to strong. As predicted, the concurrent correlation was strong between Oral Reading Fluency and Nonsense Word Fluency ($r = .73$) in the winter, while the concurrent correlation was smaller but still strong between SAT-10 and Nonsense Word Fluency ($r = .64$) in the fall. Predictive correlations were smaller than concurrent fall and winter correlations for both Oral Reading Fluency and SAT-10. Fall and winter Nonsense Word Fluency were both strongly correlated with spring Oral Reading Fluency ($r = .69$ and .68, respectively), while fall and winter Nonsense Word Fluency were both moderately correlated with spring SAT-10 ($r = .46$ and .44, respectively). The concurrent spring to spring correlation for Nonsense Word Fluency as compared to Oral Reading Fluency was strong and similar to concurrent and predictive correlations ($r = .66$). Unexpectedly, the concurrent spring to spring correlation for Nonsense Word Fluency as compared to SAT-10 was smaller than

Table 2

Overall Correlations Among All Screening and Outcome Measures

	1	2	3	4	5	6	7
1. Fall NWF-CLS	-	.64*	.74*	.76*	.57*	.69*	.46*
2. Fall SAT-10		-	.56*	.77*	.48*	.68*	.67*
3. Winter NWF-CLS			-	.73*	.72*	.68*	.44*
4. Winter ORF				-	.62*	.90*	.63*
5. Spring NWF-CLS					-	.66*	.46*
6. Spring ORF						-	.64*
7. Spring SAT-10							-

Note. NWF-CLS = Nonsense Word Fluency- Correct Letter Sounds; ORF = Oral Reading Fluency; SAT-10 = Stanford Achievement Test, 10th Edition

*$p < .01$.

fall and winter concurrent correlations and looked more similar to predictive correlations. In the spring, Nonsense Word Fluency was moderately correlated with SAT-10 (r = .46).

Research Question 2 and 2a: Overall Discriminative Ability

To answer Research Question 2 and 2a, Receiver Operator Curve (ROC) analyses were conducted to examine screeners' ability to accurately differentiate between proportions of students with and without reading difficulties on each outcome measure for the overall sample. ROC analyses were conducted for fall, winter, and spring Nonsense Word Fluency predicting to winter and spring Oral Reading Fluency and fall and spring SAT-10. For each ROC analysis, an optimal cut-score for risk was identified which prioritized the highest specificity value possible given a sensitivity value of .90 or higher (Jenkins et al., 2007). ROC curves for each analysis are depicted in Figures 2-5, and AUC values, cut scores, sensitivity, and specificity values are displayed for each ROC curve analysis in Tables 3-4.

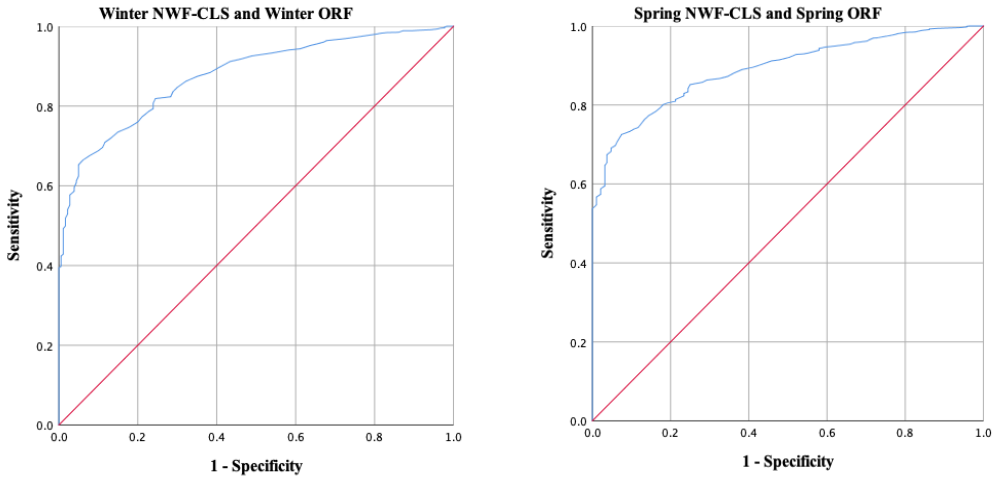
Oral Reading Fluency Discriminative Ability

Overall Accuracy. AUCs for diagnostic accuracy of Nonsense Word Fluency with concurrently administered Oral Reading Fluency in the winter (.88, 95% CI [.86, .90]) and spring (.89, 95% CI [.87, .91]) were both very good. Overlapping confidence intervals suggest that as hypothesized there was no meaningful difference in Nonsense Word Fluency's accuracy for concurrently classifying students on Oral Reading Fluency in the winter or spring.

Also as expected, non-overlapping confidence bounds around AUC values indicated that lag time meaningfully altered the accuracy of Nonsense Word Fluency for

Figure 2

ROC Curve Comparing Concurrently Administered Winter and Spring Nonsense Word Fluency (NWF-CLS) and Oral Reading Fluency (ORF)



classifying students appropriately overall, with AUC values dropping as lag time increased. The AUC for fall Nonsense Word Fluency predicting spring Oral Reading Fluency was .80, 95% CI [.77, .83], indicating a reasonable screener, while the AUC for winter Nonsense Word Fluency predicting spring Oral Reading Fluency was .87, 95% CI [.84, .89], indicating a very good screener. These findings indicated that fall Nonsense Word Fluency was meaningfully poorer at discriminating between students with and without reading difficulties on spring Oral Reading Fluency than both concurrent and predictive winter and spring Nonsense Word Fluency administrations.

Sensitivity, Specificity and Cut Scores. As predicted, specificity for Nonsense Word Fluency concurrently predicting Oral Reading Fluency in the winter and spring when sensitivity was held to at least .90 was similar and unacceptable across timepoints.

Figure 3

ROC Curve Comparing Fall, Winter and Spring Nonsense Word Fluency (NWF-CLS) Predicting Spring Oral Reading Fluency (ORF)

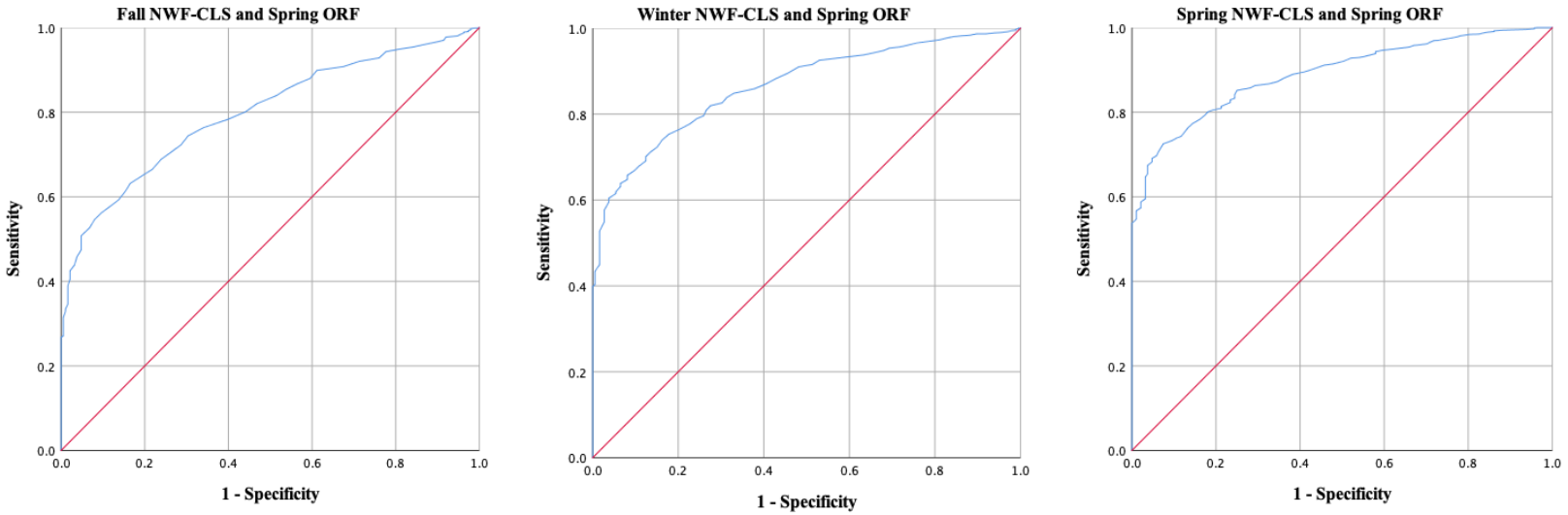


Table 3

Discriminative Ability for Nonsense Word Fluency (NWF-CLS) Predicting Oral Reading Fluency (ORF) Risk Status

NWF-CLS	Winter ORF				Spring ORF			
	AUC	Cut Score	Sens	Spec	AUC	Cut Score	Sens	Spec
				Overall				
Fall	.82 [79, .84]	28.50	.90 [.85, .94]	.41 [.38, .44]	.80 [.77, .83]	28.50	.90 [85, .94]	.39 [.36, .42]
Winter	.88 [.86, .90]	50.50	.90 [.84, .94]	.59 [.56, .62]	.87 [.84, .89]	49.50	.91 [.86, .95]	.52 [.49, .55]
Spring	-	-	-	-	.89 [.87, .91]	65.50	.90 [.85, .94]	.59 [.56, .62]
				Treatment Condition				
Fall	.83 [.79, .87]	28.50	.90 [.78, .96]	.45 [.41, .49]	.81 [.77, .85]	28.50	.90 [.81, .96]	.42 [.38, .46]
Winter	.87 [.84, .91]	52.50	.90 [.78, .96]	.57 [.53, .61]	.87 [.84, .90]	52.50	.90 [.81, .96]	.49 [.45, .53]
Spring	-	-	-	-	.87 [.84, .90]	71.50	.90 [.81, .96]	.68 [.64, .71]
				Comparison Condition				
Fall	.81 [.77, .84]	28.50	.90 [.83, .94]	.38 [.34, .42]	.79 [.75, .82]	28.50	.90 [.82, .95]	.37 [.33, .41]
Winter	.87 [.84, .90]	49.50	.90 [.83, .94]	.62 [.58, .66]	.86 [.83, .89]	47.50	.90 [.82, .95]	.53 [.49, .57]
Spring	-	-	-	-	.86 [.83, .89]	61.50	.90 [.83, .95]	.54 [.50, .58]

Note. Values in brackets indicate the 95% Confidence Interval around each diagnostic accuracy value. AUC = Area Under the Curve. Sens = Sensitivity. Spec = Specificity.

For both winter and spring Nonsense Word Fluency concurrently classifying students on winter and spring Oral Reading Fluency, a sensitivity value of .90, 95% CI [.84, .94] corresponded to a specificity value of .59, 95% CI [.56, .62]. Optimal cut scores for risk were 50.50 and 65.50 for concurrent winter and spring administrations, respectively.

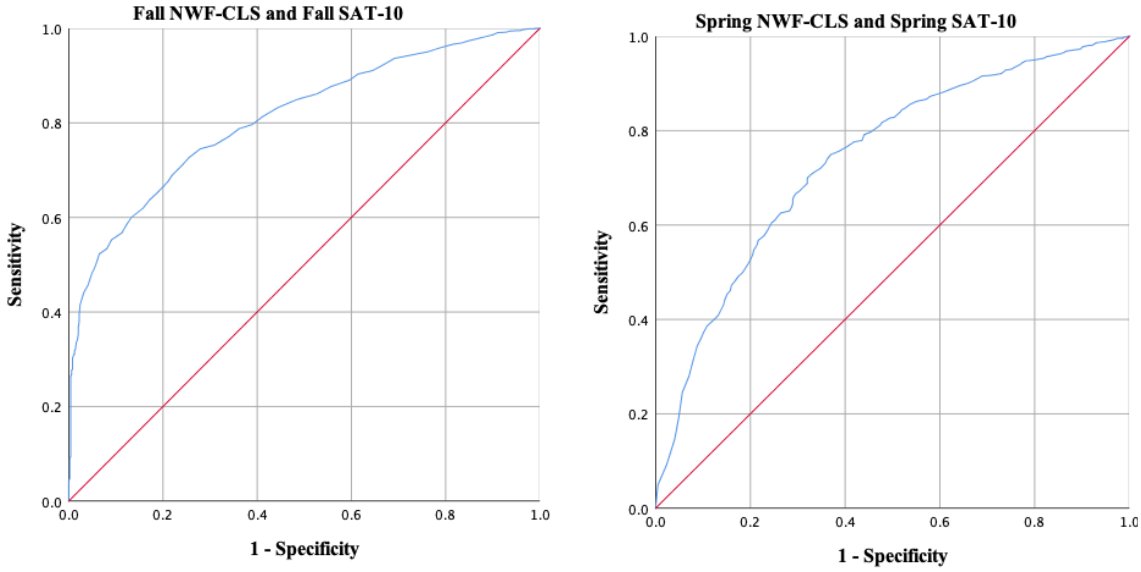
Consistent with study hypotheses, specificity values grew poorer as lag time increased. For predicting to spring Oral Reading Fluency, fall Nonsense Word Fluency had a sensitivity of .90, 95% CI [.85, .94] that corresponded to a specificity value of .39 [.36, .42], while winter Nonsense Word had a sensitivity value of .91, 95% CI [.86, .95] that corresponded to a specificity of .52, 95% CI [.49, .55]. Non-overlapping confidence intervals between all three specificity estimates provide additional evidence that increasing lag times between screener and outcome administrations resulted in meaningfully poorer Nonsense Word Fluency ability to accurately classify students who did not have reading difficulties on Oral Reading Fluency. The optimal cut score for risk for winter to spring lag time administration was 49.50, indicating that optimal cut-score for risk was one point lower when winter Nonsense Word Fluency was predicting future versus current reading difficulties.

SAT-10 Discriminative Ability

Overall Accuracy. AUCs for Nonsense Word Fluency predicting concurrently administered SAT-10 status in the fall and spring were meaningfully different in the fall as compared to the spring; the AUC for fall Nonsense Word Fluency predicting fall SAT-10 status was .81, 95% CI [.79, .83], indicating it was reasonable for decision making, while the AUC for spring Nonsense Word Fluency predicting spring SAT-10 status was .74, 95% CI [.71, .77], indicating a poor screener.

Figure 4

ROC Curve Comparing Concurrently Administered Fall and Spring Nonsense Word Fluency (NWF-CLS) and SAT-10



Across lag times, Nonsense Word Fluency was a poor screener for the purpose of predicting risk status on SAT-10. The AUC for both fall and winter Nonsense Word Fluency predicting spring SAT-10 was .73, 95% CI [.70, .76], and was not meaningfully different from the AUC for concurrently administered spring measures. Only the AUC for fall Nonsense Word Fluency predicting fall SAT-10 risk status was meaningfully different than all other time points and lag times, indicating that at this timepoint alone, Nonsense Word Fluency may be adequate for discriminative purposes.

Sensitivity, Specificity and Cut Scores. Sensitivity and specificity values were similarly poor for Nonsense Word Fluency concurrently predicting SAT-10 risk status; these values were not meaningfully different between fall and spring administrations. For

Figure 5

ROC Curve Comparing Fall, Winter and Spring Nonsense Word Fluency (NWF-CLS) Predicting Spring SAT-10

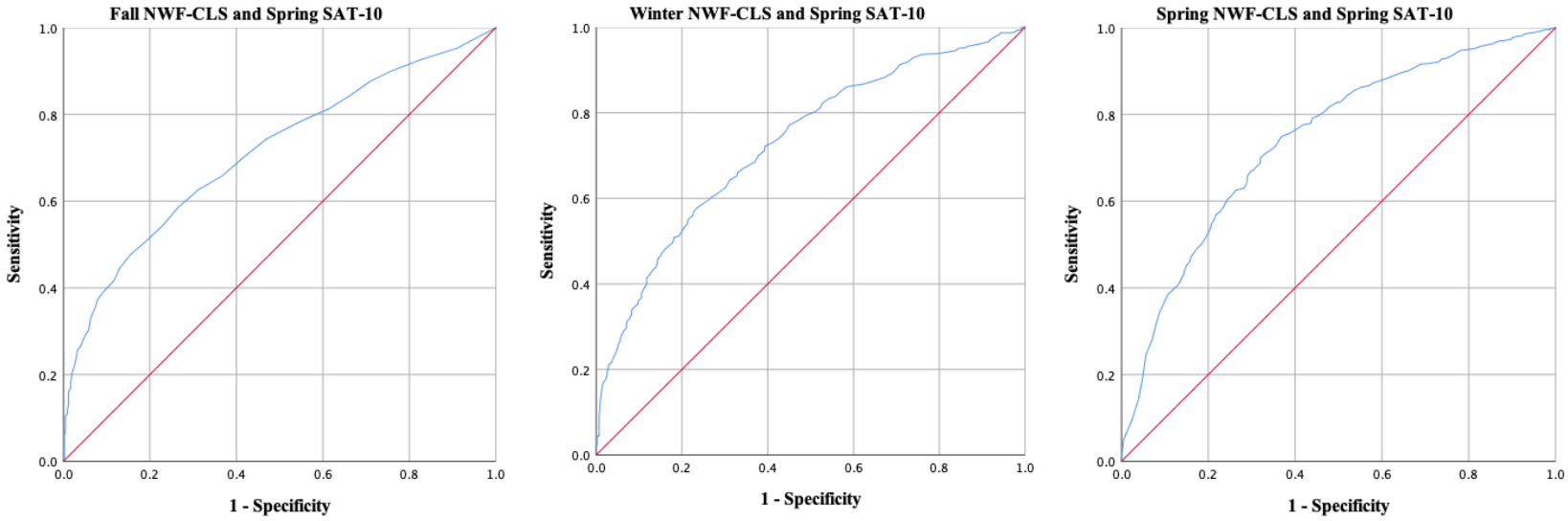


Table 4

Discriminative Ability for Nonsense Word Fluency (NWF-CLS) Predicting SAT-10 Risk Status

NWF-CLS	Fall SAT-10				Spring SAT-10			
	AUC	Cut Score	Sens	Spec	AUC	Cut Score	Sens	Spec
				Overall				
Fall	.81 [.79, .83]	31.50	.90 [.87, .93]	.39 [.36, .42]	.73 [.70, .76]	28.50	.91 [.88, .93]	.26 [.23, .29]
Winter	-	-	-	-	.73 [.70, .76]	50.50	.90 [.87, .93]	.31 [.28, .34]
Spring	-	-	-	-	.74 [.71, .77]	66.50	.90 [.87, .93]	.36 [.33, .39]
				Treatment Condition				
Fall	.83 [.80, .86]	32.50	.90 [.85, .93]	.38 [.34, .42]	.75 [.72, .79]	28.50	.91 [.86, .94]	.25 [.21, .29]
Winter	-	-	-	-	.74 [.70, .78]	55.50	.90 [.85, .94]	.36 [.32, .40]
Spring	-	-	-	-	.74 [.70, .78]	74.50	.90 [.85, .93]	.41 [.37, .45]
				Comparison Condition				
Fall	.78 [.75, .81]	31.50	.90 [.86, .93]	.37 [.33, .41]	.70 [.66, .74]	28.50	.90 [.85, .93]	.24 [.20, .28]
Winter	-	-	-	-	.72 [.68, .76]	47.50	.90 [.85, .93]	.31 [.27, .35]
Spring	-	-	-	-	.74 [.71, .78]	60.50	.90 [.85, .94]	.30 [.26, .34]

Note. Values in brackets indicate the 95% Confidence Interval around each diagnostic accuracy value. AUC = Area Under the Curve. Sens = Sensitivity. Spec = Specificity.

fall Nonsense Word Fluency predicting fall SAT-10 status, a sensitivity value of .90, 95% CI [.87, .93] corresponded to a specificity of .39, 95% CI [.36, .42]. For spring Nonsense Word Fluency predicting spring SAT-10 status, a sensitivity value of .90, 95% CI [.87, .93] corresponded to a specificity of .36, 95% CI [.33, .39]. Optimal cut scores for risk were 31.50 and 66.50 for concurrent fall and spring administrations, respectively.

As predicted, increased lag time resulted in increasingly poor specificity values. Fall Nonsense Word Fluency had a sensitivity of .91, 95% CI [.88, .93] with a specificity of .26, 95% CI [.23, .29], while winter Nonsense Word Fluency had a sensitivity of .90, 95% CI [.87, .93] with a specificity of .31, 95% CI [.28, .34]. With sensitivity held at .90 or higher, fall but not winter Nonsense Word Fluency demonstrated meaningfully poorer specificity than spring Nonsense Word Fluency for predicting spring SAT-10 risk status. This finding indicated that increased lag time resulted in meaningfully poorer Nonsense Word Fluency ability to accurately predict SAT-10 status. The optimal cut score for risk for fall to spring lag time administration was 28.50, indicating that the optimal cut-score for risk was three points lower when fall Nonsense Word Fluency was predicting future as compared to current reading difficulties.

Research Question 3 and 3a: Overall Predictive Ability

To answer Research Question 3 and 3a, positive and negative likelihood ratios and positive and negative posttest probabilities were calculated based on the sensitivity and specificity values indicated by the optimal cut-score identified for each ROC curve for the overall sample. Thus, these calculations were conducted using ROC curve analyses for fall, winter, and spring Nonsense Word Fluency predicting to winter and spring Oral Reading Fluency and fall and spring SAT-10. Positive and negative

likelihood ratios, positive and negative posttest probabilities, and base rates of reading difficulties are displayed for each ROC curve analysis in Tables 5-6.

Oral Reading Fluency Predictive Ability

Likelihood Ratios. As hypothesized, positive likelihood ratios were comparable across winter and spring concurrent administrations of Nonsense Word Fluency and Oral Reading Fluency. Positive likelihood ratios were 2.20, 95% CI [2.02, 2.38], indicating a small increase in likelihood of reading difficulties. Thus, Nonsense Word Fluency in the current study appeared borderline reasonable for the purpose of low stakes decision making based on predictions about individual students' likelihood of having current reading difficulties given an "at risk" screening result.

Positive likelihood ratios grew meaningfully poorer with increased lag times. For fall Nonsense Word Fluency predicting spring Oral Reading Fluency, the positive likelihood ratio was 1.48, 95% CI [1.38, 1.58], while the positive likelihood ratio for winter Nonsense Word Fluency predicting spring Oral Reading Fluency was 1.90, 95% CI [1.76, 2.04]. Fall, but not winter, Nonsense Word Fluency demonstrated a meaningfully poorer positive likelihood ratio than spring Nonsense Word Fluency for predicting spring Oral Reading Fluency risk status. Both fall and winter Nonsense Word Fluency likelihood ratios corresponded to a minimal increase in likelihood of having reading difficulties with a positive test result. These findings indicated that particularly in the fall, Nonsense Word Fluency was not an appropriate tool for determining an individual student's likelihood of having reading difficulties on Oral Reading Fluency in the future given an "at risk" screening result within the current study context.

Negative likelihood ratios for both winter and spring concurrent administrations of Nonsense Word Fluency and Oral Reading Fluency were 0.17, 95% CI [.17, .26], indicating a moderate decrease in likelihood of reading difficulties. Similarly, negative likelihood ratios were .26, 95% CI [.17, .40] for fall Nonsense Word Fluency and .17, 95% CI [.11, .28] for winter Nonsense Word Fluency predicting spring Oral Reading Fluency risk status, indicating small and moderate decreases in likelihood of reading difficulties with a negative test result for fall and winter, respectively. These values were not meaningfully different than the negative likelihood ratio for spring Nonsense Word Fluency predicting spring Oral Reading Fluency risk status, suggesting that across the year, Nonsense Word Fluency was a borderline appropriate tool for making low stakes decisions based on individual students' likelihood of having reading difficulties given a "not at risk" screening result.

Posttest Probabilities and Base Rates. Contrary to study hypotheses, winter and spring posttest probabilities were nearly identical for concurrently administered Nonsense Word Fluency and Oral Reading Fluency, with positive posttest probabilities of 23% and 25% for winter and spring calculations, respectively, and negative posttest probabilities of 2% for both winter and spring calculations. These similar statistics may be partially attributable to similar base rates of reading difficulties according to Oral Reading Fluency in both winter (12%) and spring (13%). Based on recommendations by VanDerHeyden (2013), when concurrently administered with Oral Reading Fluency in the winter and spring, Nonsense Word Fluency was appropriate for ruling out reading difficulties but was insufficient for ruling in reading difficulties for the current sample.

Posttest probabilities were also not substantially different across lag times. Positive posttest probabilities were 18% for fall Nonsense Word Fluency and 22% for winter Nonsense Word Fluency, while negative posttest probabilities were 4% for fall Nonsense Word Fluency and 2% for winter Nonsense Word Fluency, again suggesting that across lag times, Nonsense Word Fluency may be adequate for ruling out reading difficulties and inadequate for ruling in reading difficulties for students in the current sample.

SAT-10 Predictive Ability

Likelihood Ratios. Similar to Oral Reading Fluency, positive and negative likelihood ratios were not meaningfully different for concurrently administered Nonsense Word Fluency and SAT-10 at each time point. Positive likelihood ratios were 1.48, 95% CI [1.39, 1.56] and 1.41, 95% CI [1.33, 1.49] for fall and spring, respectively, indicating no meaningful change in likelihood of having reading difficulties with a positive test result. This finding suggested that Nonsense Word Fluency may not be appropriate for predicting an individual student's likelihood of current reading difficulties on SAT-10 given an "at risk" screening result.

As expected, increased lag time resulted in increasingly poor positive likelihood ratios for Nonsense Word Fluency predicting spring SAT-10 risk. Positive likelihood ratios were 1.23, 95% CI [1.17, 1.29] and 1.30, 95% CI [1.24, 1.37] for fall and winter, respectively, indicating no meaningful increase in likelihood of reading difficulties given an "at risk" screening result. Non-overlapping confidence intervals indicated that fall, but not winter, Nonsense Word Fluency demonstrated a meaningfully poorer positive likelihood ratio than spring Nonsense Word Fluency. Overall, these findings indicated

Table 5

Predictive Ability for Nonsense Word Fluency (NWF-CLS) Predicting Oral Reading Fluency (ORF) Risk Status

NWF-CLS	Winter ORF					Spring ORF				
	Pos LR	Neg LR	PT Prob+	PT Prob-	Base Rate	Pos LR	Neg LR	PT Prob+	PT Prob-	Base Rate
					Overall					
Fall	1.53 [1.43, 1.63]	0.24 [0.16, 0.38]	17%	3%	12%	1.48 [1.38, 1.58]	0.26 [0.17, 0.40]	18%	4%	13%
Winter	2.20 [2.02, 2.38]	0.17 [0.11, 0.26]	23%	2%	12%	1.90 [1.76, 2.04]	0.17 [0.11, 0.28]	22%	2%	13%
Spring	-	-	-	-	-	2.20 [2.03, 2.38]	0.17 [0.11, 0.26]	25%	2%	13%
					Treatment Condition					
Fall	1.64 [1.46, 1.83]	0.22 [0.10, 0.49]	12%	2%	8%	1.55 [1.40, 1.72]	0.24 [0.12, 0.47]	16%	3%	11%
Winter	2.09 [1.85, 2.37]	0.18 [0.08, 0.38]	15%	2%	8%	1.76 [1.58, 1.97]	0.20 [0.10, 0.40]	18%	2%	11%
Spring	-	-	-	-	-	2.81 [2.46, 3.22]	0.15 [0.07, 0.29]	24%	2%	10%
					Comparison Condition					
Fall	1.45 [1.33, 1.58]	0.26 [0.15, 0.45]	22%	5%	16%	1.43 [1.31, 1.56]	0.27 [0.15, 0.48]	20%	5%	15%
Winter	2.37 [2.11, 2.66]	0.16 [0.09, 0.28]	31%	3%	16%	1.91 [1.72, 2.13]	0.19 [0.11, 0.33]	25%	3%	15%
Spring	-	-	-	-	-	1.96 [1.76, 2.17]	0.19 [0.11, 0.32]	26%	3%	15%

Note. Values in brackets indicate a 95% Confidence Interval around each diagnostic accuracy value. Pos LR = Positive likelihood ratio. Neg LR = Negative likelihood ratio. PT Prob+ = Posttest probability of true reading difficulty for a positive test result. PT Prob- = Posttest probability of true reading difficulty for a negative test result.

that across timepoints and lag time, Nonsense Word Fluency was not adequate for ruling in reading difficulties on spring SAT-10.

Negative likelihood ratios were slightly more useful for decision making. Negative likelihood ratios were .26, 95% CI [.19, .34] and .28, 95% CI [.21, .37] for concurrent fall and spring administrations, respectively, indicating a small decrease in likelihood of currently having reading difficulties with a "not at risk" screening result. These values were slightly, though not meaningfully, poorer for lag time administrations. Negative likelihood ratios were .35, 95% CI [.25, .48] and .32, 95% CI [.24, .43] for fall and winter Nonsense Word Fluency predicting spring SAT-10, respectively, indicating a small decrease in likelihood of future reading difficulties with a "not at risk" test result. Thus, across the year, Nonsense Word Fluency appeared adequate for making low stakes decisions based around ruling out future reading difficulties for certain students given a "not at risk" screening result.

Posttest Probabilities and Base Rates. Base rate of reading difficulties on SAT-10 for the overall sample were similar across fall (33%) and spring (28%). This resulted in an expected small decrease in positive posttest probabilities across the year from 42% in the fall to 35% in the spring. At both timepoints an "at risk" screening result on Nonsense Word Fluency indicated only a slight increase in likelihood of current reading difficulties, and not nearly enough of an increase to meet VanDerHeyden (2013)'s threshold for a decision to provide intervention. Unexpectedly, negative posttest probabilities remained similar across the year, with negative posttest probabilities of 11% in the fall and 10% in the spring. These values indicate that in both the fall and the spring, Nonsense Word Fluency did an inadequate job of ruling in current reading difficulties

Table 6

Predictive Ability for Nonsense Word Fluency (NWF-CLS) Predicting SAT-10 Risk Status

NWF-CLS	Fall SAT-10					Spring SAT-10				
	Pos LR	Neg LR	PT Prob+	PT Prob-	Base Rate	Pos LR	Neg LR	PT Prob+	PT Prob-	Base Rate
					Overall					
Fall	1.48 [1.39, 1.56]	0.26 [0.19, 0.34]	42%	11%	33%	1.23 [1.17, 1.29]	0.35 [0.25, 0.48]	35%	13%	30%
Winter	-	-	-	-	-	1.30 [1.24, 1.37]	0.32 [0.24, 0.43]	36%	12%	30%
Spring	-	-	-	-	-	1.41 [1.33, 1.49]	0.28 [0.21, 0.37]	35%	10%	28%
					Treatment Condition					
Fall	1.48 [1.39, 1.56]	0.26 [0.19, 0.34]	42%	11%	33%	1.21 [1.13, 1.30]	0.36 [0.23, 0.57]	33%	13%	29%
Winter	-	-	-	-	-	1.41 [1.30, 1.52]	0.28 [0.18, 0.43]	37%	10%	29%
Spring	-	-	-	-	-	1.53 [1.40, 1.66]	0.24 [0.16, 0.37]	37%	9%	28%
					Comparison Condition					
Fall	1.48 [1.39, 1.56]	0.26 [0.19, 0.34]	42%	11%	33%	1.18 [1.11, 1.26]	0.42 [0.27, 0.63]	35%	16%	31%
Winter	-	-	-	-	-	1.30 [1.21, 1.40]	0.32 [0.21, 0.49]	36%	12%	30%
Spring	-	-	-	-	-	1.29 [1.20, 1.38]	0.33 [0.22, 0.51]	35%	12%	29%

Note. Values in brackets indicate a 95% Confidence Interval around each diagnostic accuracy value. Pos LR = Positive likelihood ratio. Neg LR = Negative likelihood ratio. PT Prob+ = Posttest probability of true reading difficulty for a positive test result. PT Prob- = Posttest probability of true reading difficulty for a negative test result.

and did a borderline adequate job of ruling out current reading difficulties in the current sample.

The base rates of reading difficulties for fall and winter Nonsense Word Fluency predicting spring SAT-10 risk status was 30%, resulting in positive posttest probabilities of 35% and 36% and negative posttest probabilities of 13% and 12% for fall and winter, respectively. Thus, posttest probabilities were similar for predicting current versus future reading difficulties, and in both instances, Nonsense Word Fluency was generally inadequate for the purpose of ruling in or ruling out reading difficulties.

Research Question 4: Validity by Instructional Effectiveness

To answer Research Question 4, Pearson's *r* correlations and Receiver Operator Curve (ROC) analyses were examined for each combination of screener and outcome measures at each time point and lag time separately for the treatment (i.e., higher instructional effectiveness) and comparison (i.e., lower instructional effectiveness) conditions. Correlation coefficients for the treatment and comparison condition are displayed in Tables 7-8. ROC curves comparing treatment and comparison condition analyses are depicted in Figures 6-13. AUC values, optimal cut-scores, sensitivity, and specificity for each analysis broken down by treatment (ECRI) and comparison (Control) condition are displayed in Tables 3-4. Positive and negative likelihood ratios, posttest probabilities and base rates are reported by condition in Tables 5-6.

Test Score Interpretations

Concurrent correlations between Nonsense Word Fluency and both outcome measures were similar across conditions. For Nonsense Word Fluency and Oral Reading Fluency, concurrent winter correlations were .70 in the treatment condition and .75 in the

comparison condition, while concurrent spring correlations were .63 in the treatment condition and .68 in the comparison condition. For Nonsense Word Fluency and SAT-10, concurrent fall correlations were .64 in the treatment condition, and .63 in the comparison condition, while concurrent spring correlations were .46 in the treatment condition and .47 in the comparison condition. Thus, regardless of instructional effectiveness condition, in the current study concurrent fall correlations indicated strong evidence and concurrent spring correlations indicated moderate to strong evidence for Nonsense Word Fluency scores as indicative of students' skills in oral reading fluency and overall reading achievement.

Predictive correlations between Nonsense Word Fluency and Oral Reading Fluency demonstrated only minor differences between conditions across lag times, with the treatment condition demonstrating slightly smaller winter to spring correlations. Fall to spring correlations were .67 in the treatment condition and .70 in the comparison condition, while winter to spring correlations were .64 in the treatment condition and .71 in the comparison condition. Predictive correlations demonstrated small differences in the opposite direction on SAT-10, with slightly weaker values in the comparison condition than the treatment condition from fall to spring. Fall to spring correlations were .51 in the treatment condition and .43 in the comparison condition. Winter to spring correlations were .45 in the treatment condition and .43 in the comparison condition. Regardless, across instructional effectiveness conditions and lag times, predictive correlations provided evidence that Nonsense Word Fluency was strongly indicative of students' oral reading fluency skills and moderately indicative of students' overall reading achievement.

Table 7

Treatment Condition Correlations Among All Screening and Outcome Measures

	1	2	3	4	5	6	7
1. Fall NWF-CLS	-	.64**	.74**	.76**	.56**	.67**	.51**
2. Fall SAT-10		-	.55**	.76**	.45**	.65**	.66**
3. Winter NWF-CLS			-	.70**	.69**	.64**	.45**
4. Winter ORF				-	.60**	.89**	.65**
5. Spring NWF-CLS					-	.63**	.46**
6. Spring ORF						-	.63**
7. Spring SAT-10							-

**$p < .01$.

Table 8

Control Condition Correlations Among All Screening and Outcome Measures

	1	2	3	4	5	6	7
1. Fall NWF-CLS	-	.63**	.74**	.77**	.59**	.70**	.43**
2. Fall SAT-10		-	.57**	.77**	.50**	.71**	.69**
3. Winter NWF-CLS			-	.75**	.74**	.71**	.43**
4. Winter ORF				-	.63**	.91**	.62**
5. Spring NWF-CLS					-	.68**	.47**
6. Spring ORF						-	.65**
7. Spring SAT-10							-

**$p < .01$.

Test Score Uses: Discriminative Ability

Oral Reading Fluency Overall Accuracy. AUC values for Nonsense Word Fluency predicting Oral Reading Fluency risk status were similar across treatment and comparison conditions at all time points and lag times, with overlapping confidence intervals. For concurrent winter administration, the AUC value was .87 for both conditions, with 95% CI [.84, .91] and 95% CI [.84, .90] for treatment and comparison conditions, respectively. For concurrent spring administration, the AUC value was .90, 95% CI [.87, .93] for the treatment condition and .88, 95% CI [.85, .91] for the comparison condition. Thus, as expected, AUC values indicated that regardless of instructional context, Nonsense Word Fluency was a very good tool for accurately discriminating between students with and without current reading difficulties on Oral Reading Fluency.

Unexpectedly, lag time administrations did not result in meaningfully different diagnostic accuracy across instructional contexts. AUC values were .81, 95% CI [.77, .85] and .79, 95% CI [.75, .82] for fall Nonsense Word Fluency predicting spring Oral Reading Fluency risk status in the treatment and comparison conditions, respectively. These values demonstrated that across instructional contexts fall Nonsense Word Fluency was a reasonable tool for discriminating between students with and without reading difficulties on spring Oral Reading Fluency. AUC values were .87, 95% CI [.84, .90] and .86, 95% CI [.83, .89] for winter Nonsense Word Fluency predicting spring Oral Reading Fluency risk status in the treatment and comparison conditions, respectively, indicating that winter Nonsense Word Fluency was a very good tool for discriminating between students with and without reading difficulties on spring Oral Reading Fluency.

Figure 6

ROC Curve Comparing Treatment vs. Comparison Condition Winter Nonsense Word Fluency (NWF-CLS) Predicting Winter Oral Reading Fluency (ORF) Risk Status

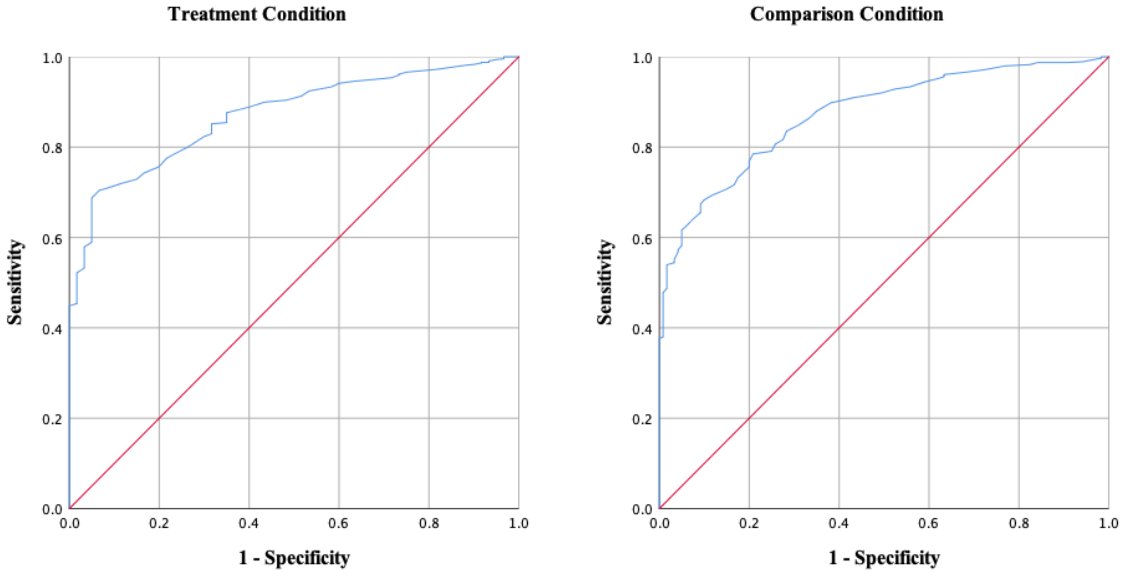


Figure 7

ROC Curve Comparing Treatment vs. Comparison Condition Spring Nonsense Word Fluency (NWF-CLS) Predicting Spring Oral Reading Fluency (ORF) Risk Status

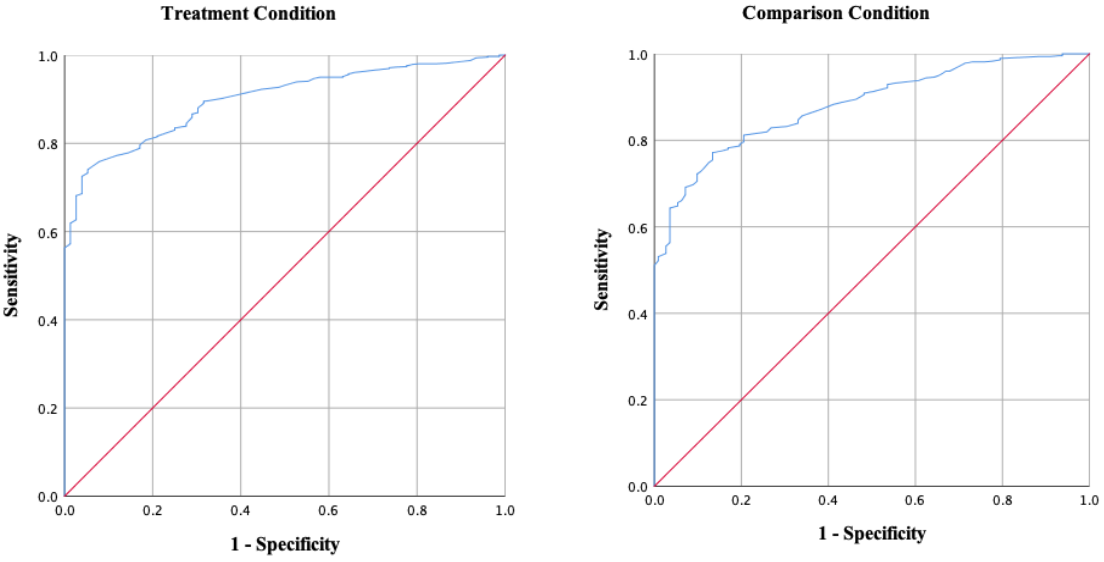


Figure 8

ROC Curve Comparing Treatment vs. Comparison Condition Fall Nonsense Word Fluency (NWF-CLS) Predicting Spring Oral Reading Fluency (ORF) Risk Status

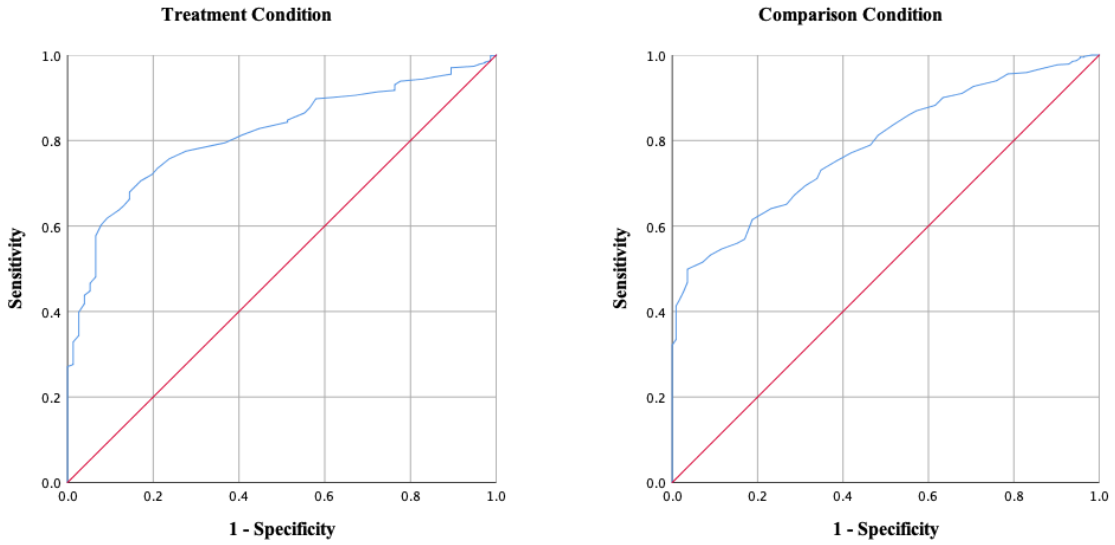
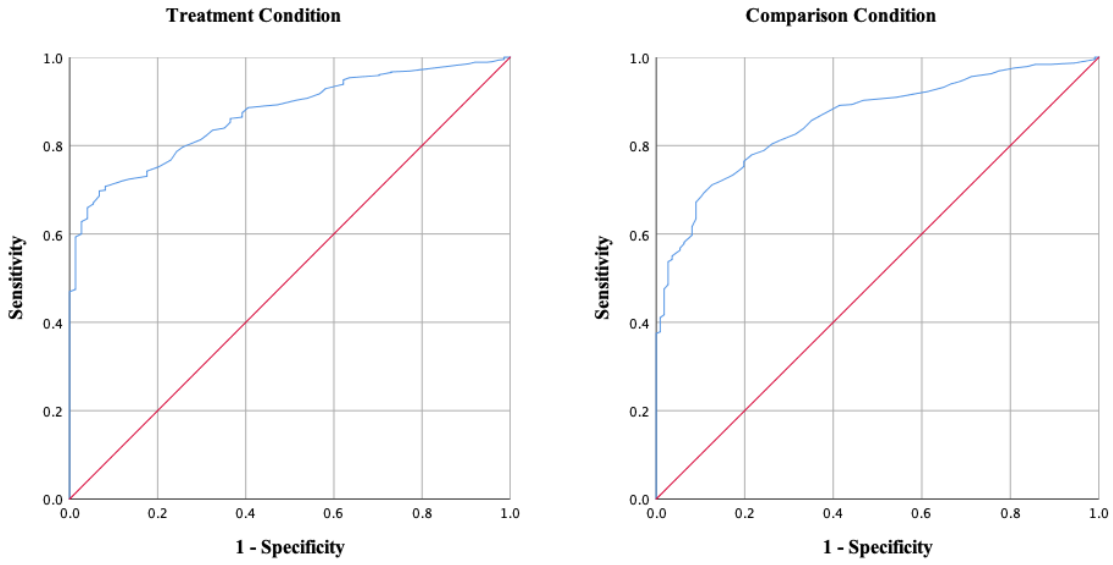


Figure 9

ROC Curve Comparing Treatment vs. Comparison Condition Winter Nonsense Word Fluency (NWF-CLS) Predicting Spring Oral Reading Fluency (ORF) Risk Status



Oral Reading Fluency Sensitivity, Specificity and Cut Scores. In most cases, optimal cut scores were similar, with sensitivity and specificity values with overlapping confidence intervals across instructional contexts. For winter Nonsense Word Fluency predicting winter Oral Reading Fluency status, in the treatment condition, the optimal cut score was 52.50 with a sensitivity of .90, 95% CI [.78, .96] corresponding to a specificity of .57, 95% CI [.53, .61], while in the comparison condition, the optimal cut score was 49.50 with a sensitivity of .90, 95% CI [.83, .94] corresponding to a specificity of .62, 95% CI [.58, .66]. Thus, in a setting with stronger instructional effectiveness, students could earn a score on winter Nonsense Word Fluency up to 3 points higher than in a setting with lower instructional effectiveness and still be identified as at risk for reading difficulties on winter Oral Reading Fluency.

Lag time administrations resulted in meaningful differences in diagnostic accuracy between instructional contexts in some, but not all, cases. For fall Nonsense Word Fluency predicting spring Oral Reading Fluency status, the optimal cut score for risk was 28.50 in both conditions. This corresponded to a sensitivity value of .90, 95% CI [.81, .96] and specificity value of .42, 95% CI [.38, .46] in the treatment condition, and a sensitivity value of .90, 95% CI [.82, .95] with a specificity value of .37, 95% CI [.33, .41] in the comparison condition.

For winter Nonsense Word Fluency predicting spring Oral Reading Fluency status, in the treatment condition the optimal cut score for risk was 52.50 with a sensitivity value of .90, 95% CI [.81, .96] corresponding to a specificity value of .49, 95% CI [.45, .53], while in the comparison condition the optimal cut score for risk was 47.50 with a sensitivity of .90, 95% CI [.81, .95] corresponding to a specificity value of

.53, 95% CI [.49, .57]. Thus, though specificity values did not meaningfully vary, students in the higher instructional effectiveness condition who received a winter Nonsense Word Fluency score up to five points higher than students in the lower instructional effectiveness condition were classified as having reading difficulties on spring Oral Reading Fluency.

Concurrent spring diagnostic accuracy for Nonsense Word Fluency predicting Oral Reading Fluency indicated widely varying cut scores for risk and stronger specificity in the treatment condition when holding sensitivity to .90. Specifically, in the treatment condition, the optimal cut score for risk was 71.50 and corresponded to a sensitivity value of .90, 95% CI [.81, .96] and specificity value of .68, 95% CI [.64, .71]. In contrast, in the comparison condition, the optimal cut score for risk was 61.50 and corresponded to a sensitivity value of .90, 95% CI [.83, .95] and specificity value of .54, 95% CI [.50, .58]. These findings suggested that in the spring, Nonsense Word Fluency was stronger at accurately identifying students who were not at risk in a context with stronger instructional effectiveness, and that the cut score that represented the optimal balance of sensitivity and specificity varied by ten points based on instructional context. Students in the more effective instructional setting needed to receive a Nonsense Word Fluency score that was 10 points higher than students in the lower instructional effectiveness condition to no longer be classified as having reading difficulties on spring Oral Reading Fluency.

SAT-10 Overall Accuracy. Across timepoints and lag times, AUCs were similar with overlapping confidence intervals across conditions. For fall Nonsense Word Fluency predicting fall SAT-10, AUC values were .83, 95% CI [.80, .86] and .78, 95% CI [.75,

Figure 10

ROC Curve Comparing Treatment vs. Comparison Condition Fall Nonsense Word Fluency (NWF-CLS) Predicting Fall SAT-10 Risk Status

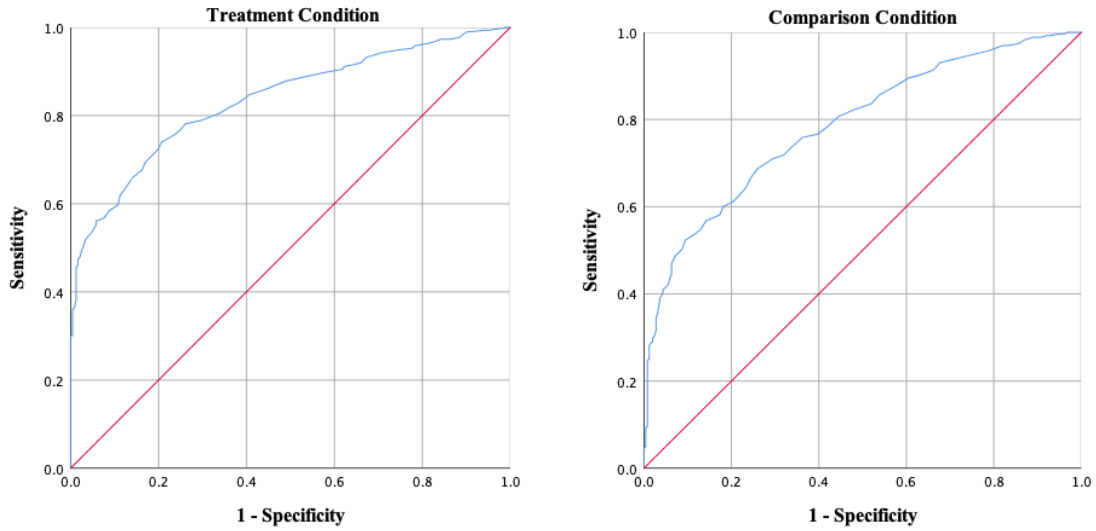


Figure 11

ROC Curve Comparing Treatment vs. Comparison Condition Spring Nonsense Word Fluency (NWF-CLS) Predicting Spring SAT-10 Risk Status

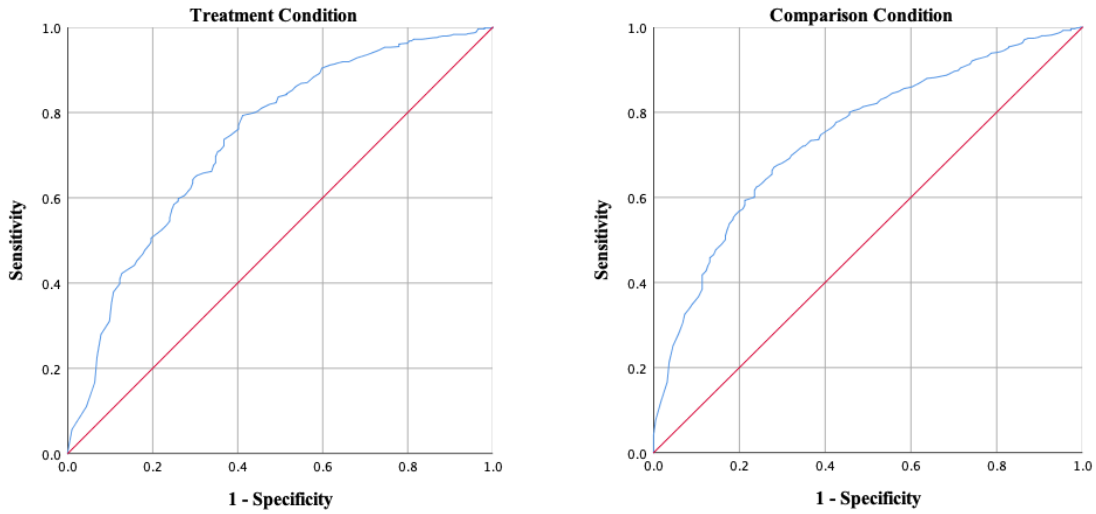


Figure 12

ROC Curve Comparing Treatment vs. Comparison Condition Fall Nonsense Word Fluency (NWF-CLS) Predicting Spring SAT-10 Risk Status

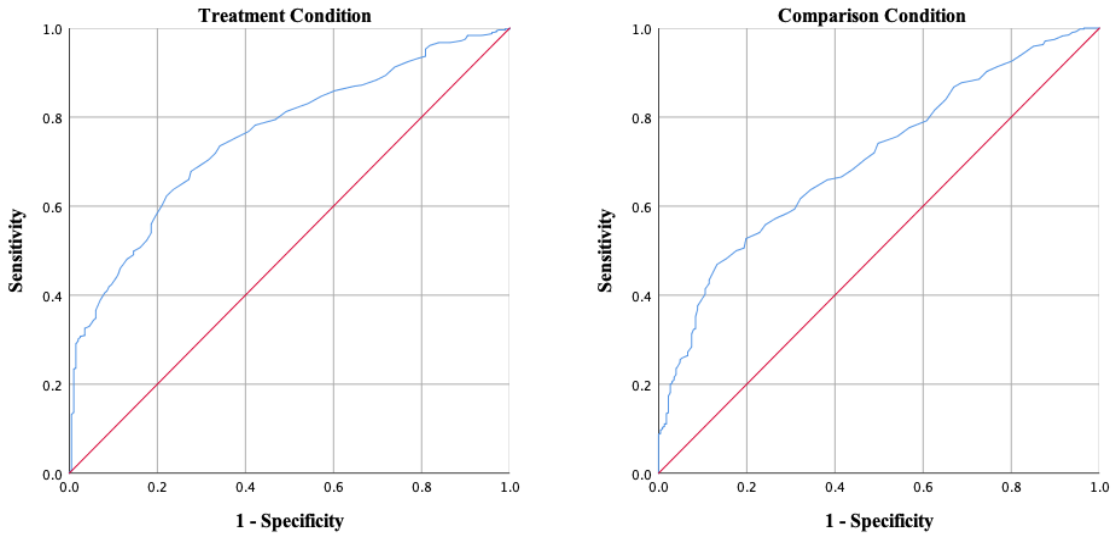
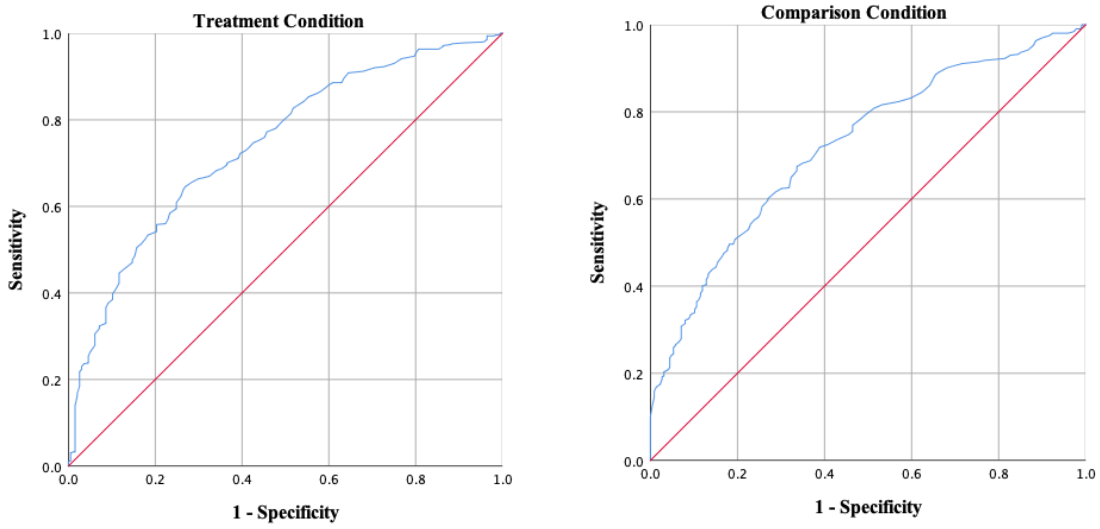


Figure 13

ROC Curve Comparing Treatment vs. Comparison Condition Winter Nonsense Word Fluency (NWF-CLS) Predicting Spring SAT-10 Risk Status



.81] for treatment and comparison conditions, respectively. For spring Nonsense Word Fluency predicting spring SAT-10, AUC values were .74, 95% CI [.70, .78] and .74, 95% CI [.71, .78] for treatment and comparison conditions, respectively. Thus, as expected, concurrent administrations resulted in similar overall diagnostic accuracy regardless of instructional context.

Unexpectedly, predictive diagnostic accuracy followed a similar pattern. Fall Nonsense Word Fluency predicting spring SAT-10 had an AUC of .75, 95% CI [.72, .79] and .70, 95% CI [.66, .74] in the treatment and comparison conditions, respectively, while winter Nonsense Word Fluency predicting spring SAT-10 had an AUC of .74, 95% CI [.70, .78] and .72, 95% CI [.68, .76] in the treatment and comparison conditions, respectively. Thus, across timepoints and time lags, Nonsense Word Fluency demonstrated similarly poor discriminative ability regardless of instructional effectiveness.

SAT-10 Sensitivity, Specificity and Cut Scores. Cut scores, sensitivity, and specificity values were similar for concurrent fall administrations between conditions. For fall Nonsense Word Fluency predicting fall SAT-10, the optimal cut score for risk in the treatment condition was 32.50, which corresponded to a sensitivity value of .90, 95% CI [.85, .93] and a specificity value of .38, 95% CI [.34, .42]. In the comparison condition, the optimal cut score for risk was 31.50, which corresponded to a sensitivity value of .90, 95% CI [.86, .93] and a specificity value of .37, 95% CI [.33, .41]. Thus, optimal cut scores for risk varied by just one point between conditions.

Similar to the findings for Oral Reading Fluency, sensitivity and specificity values for fall and winter Nonsense Word Fluency predicting spring SAT-10 were also

comparable between conditions, but with more variable cut scores. For an optimal cut score of 28.50 across conditions, sensitivity values ranged from .90 to .91 with overlapping confidence intervals, and specificity values were .25, 95% CI [.21, .29] and .24, 95% CI [.20, .28] in the treatment and comparison conditions, respectively for fall Nonsense Word Fluency. Optimal cut scores for risk varied for winter Nonsense Word Fluency predicting spring SAT-10, with a cut score of 55.50 with a specificity of .36, 95% CI [.32, .40] for a sensitivity of .90 in the treatment condition and a cut score of 47.50 with a specificity of .31, 95% CI [.27, .35] for a sensitivity of .90 in the comparison condition. These values demonstrated that regardless of condition, Nonsense Word Fluency did a similarly poor job of classifying students who were actually not at risk on spring SAT-10 when sensitivity was maximized. At the same time, the optimal cut score for risk differed by eight points based on instructional context, indicating that in the higher instructional effectiveness condition, a substantially higher score on winter Nonsense Word Fluency was needed to be classified as having no reading difficulties on spring SAT-10 than in the lower instructional effectiveness condition.

Similar to Oral Reading Fluency results and contrary to study predictions, concurrent spring administrations resulted in widely varying cut scores and specificity values that were meaningfully stronger in the treatment condition for predicting SAT-10 performance. In the treatment condition, the optimal cut score for risk was 74.50, which corresponded to a sensitivity value of .90, 95% CI [.85, .93] and a specificity value of .41, 95% CI [.37, .45]. In the comparison condition, the optimal cut score for risk was 60.50, which corresponded to a sensitivity value of .90, 95% CI [.85, .94] and a specificity value of .30, 95% CI [.26, .34]. These findings demonstrated that in the spring

only, Nonsense Word Fluency did a meaningfully better job of accurately classifying students who were not at risk in a more effective instructional context. Additionally, optimal cut scores for risk varied widely depending on instructional context, with a spring Nonsense Word Fluency score that was 14 points higher being necessary to be classified as not at risk on spring SAT-10 in the higher instructional effectiveness condition than in the lower instructional effectiveness condition.

Test Score Uses: Predictive Ability

Oral Reading Fluency Likelihood Ratios. Similar to discriminative ability findings, positive and negative likelihood ratios were not meaningfully different between conditions across concurrent winter and predictive winter and spring administrations. For fall Nonsense Word Fluency predicting fall Oral Reading Fluency, positive likelihood ratios were 2.09, 95% CI [1.85, 2.37] and 2.37, 95% CI [2.11, 2.66] and negative likelihood ratios were 0.18, 95% CI [.08, .38] and 0.16, 95% CI [.09, .28] in the treatment and comparison conditions, respectively. For fall Nonsense Word Fluency predicting spring Oral Reading Fluency risk status, positive likelihood ratios were 1.55, 95% CI [1.40, 1.72] and 1.43, 95% CI [1.31, 1.56] in the treatment and comparison conditions, respectively, while negative likelihood ratios were 0.24, 95% CI [0.12, 0.47] and 0.27, 95% CI [0.15, 0.48] in the treatment and comparison conditions, respectively. For winter Nonsense Word Fluency predicting spring Oral Reading Fluency risk status, positive likelihood ratios were 1.76, 95% CI [1.58, 1.97] and 1.91, 95% CI [1.72, 2.13] for treatment and comparison conditions, respectively, while negative likelihood ratios were 0.20, 95% CI [0.10, .0.40] and 0.19, 95% CI [0.11, 0.33] for treatment and comparison conditions, respectively.

Again, concurrent spring administration was the one exception, with meaningfully stronger positive likelihood ratios in the treatment condition. Positive likelihood ratios were 2.81, 95% CI [2.46, 3.22] and 1.96, 95% CI [1.76, 2.17] in the treatment and comparison conditions respectively, suggesting that in the treatment condition, a positive test result indicated a small increase in likelihood of reading difficulties while in the comparison condition a positive test result did not indicate any meaningful change in likelihood of having reading difficulties on Oral Reading Fluency. In contrast, negative likelihood ratios were similar across conditions, with values of .15, 95% CI [.07, .29] and .19, 95% CI [.11, .32] in the treatment and comparison conditions, respectively. Thus, it appeared that only in the case of spring Nonsense Word Fluency predicting the likelihood of current reading difficulties on Oral Reading Fluency, Nonsense Word Fluency was more useful for ruling in reading difficulties in a more effective instructional condition.

Oral Reading Fluency Posttest Probabilities and Base Rates. Posttest probabilities in the current study were similar between instructional contexts for lag time administrations. Based on spring base rates of reading difficulties of 11% and 15% on Oral Reading Fluency in the treatment and comparison conditions, respectively, fall Nonsense Word Fluency produced positive posttest probabilities of 16% and 20%, and negative posttest probabilities of 3% and 5% in the treatment and comparison conditions, respectively. Similarly, winter Nonsense Word Fluency produced positive posttest probabilities of 18% and 25%, and negative posttest probabilities of 2% and 3% in the treatment and comparison conditions, respectively.

Concurrent administrations demonstrated more varied posttest probabilities despite similar population-based diagnostic accuracy statistics (i.e., AUCs, sensitivity,

specificity, likelihood ratios) between instructional contexts. The base rate of reading difficulties as defined by winter Oral Reading Fluency was 8% in the treatment condition and 16% in the comparison condition. For winter Nonsense Word Fluency predicting winter Oral Reading Fluency, positive posttest probabilities were 15% in the treatment condition and 31% in the comparison condition, and negative posttest probabilities were 2% in the treatment condition and 3% in the comparison condition.

In contrast, despite meaningfully different specificity values between conditions for spring Nonsense Word Fluency concurrently predicting spring Oral Reading Fluency outcomes, posttest probabilities were nearly the same. Positive posttest probabilities were 24% and 26% and negative posttest probabilities were 2% and 3% for treatment and comparison conditions, respectively. These findings demonstrate that in certain contexts posttest probabilities may vary based on instructional context. They also show that overall, Nonsense Word Fluency did an adequate job of ruling out but not ruling in current and future reading difficulties on Oral Reading Fluency across instructional conditions.

SAT-10 Likelihood Ratios. In most cases there were no meaningful differences in likelihood ratios for Nonsense Word Fluency predicting SAT-10 between treatment and comparison conditions. For fall Nonsense Word Fluency predicting fall SAT-10 risk status, positive likelihood ratios were 1.45, 95% CI [1.34, 1.57] and 1.43, 95% CI [1.32, 1.54], while negative likelihood ratios were 0.26, 95% CI [0.18, 0.39] and 0.27, 95% CI [0.18, 0.40] for treatment and comparison conditions, respectively in the fall. Thus, as expected, there were no meaningful differences in likelihood ratios across conditions in the fall.

Contrary to study hypotheses, there were also no meaningful differences between conditions for fall and winter Nonsense Word Fluency predicting spring SAT-10 risk status. Positive likelihood ratios were 1.45, 95% CI [1.34, 1.57] and 1.43, 95% CI [1.32, 1.54] in the fall, and 1.41, 95% CI [1.30, 1.52] and 1.30, 95% CI [1.21, 1.40] in the winter for treatment and comparison conditions, respectively. Negative likelihood ratios were 0.36, 95% CI [0.23, 0.57] and 0.42, 95% CI [0.27, 0.63] in the fall and 0.28, 95% CI [0.18, 0.43] and 0.32, 95% CI [0.21, 0.49] in the winter, for treatment and comparison conditions, respectively.

Similar to Oral Reading Fluency, spring Nonsense Word Fluency predicting spring SAT-10 risk status was the exception. In this case, the positive likelihood ratio was meaningfully stronger in the treatment condition, suggesting that in the spring only, Nonsense Word Fluency did a better job of ruling in reading difficulties than in the comparison condition. Specifically, positive likelihood ratios were 1.53, 95% CI [1.40. 1.66] in the treatment condition as compared to 1.29, 95% CI [1.20, 1.38] in the comparison condition. Negative likelihood ratios were similar with overlapping confidence intervals across conditions: negative likelihood ratios were 0.24, 95% CI [0.16, 0.37] in the treatment condition and 0.33, 95% CI [0.22, 0.51] in the comparison condition.

SAT-10 Posttest Probabilities and Base Rates. In both conditions, the base rate of reading difficulties on fall SAT-10 was 33%. This corresponded to similar posttest probabilities across conditions for concurrent administrations, as expected. Positive posttest probabilities were 42% and 41%, and negative posttest probabilities were 11% and 12% for treatment and comparison conditions, respectively for concurrently

administered fall assessments. However, contrary to study hypotheses, posttest probabilities also did not substantially differ for concurrent spring administrations based on instructional context. For concurrent spring administrations, positive posttest probabilities were 37% and 35%, and negative posttest probabilities were 9% and 12% across treatment and comparison conditions, respectively.

Posttest probabilities for lag time administrations similarly did not differ substantially across conditions. Positive posttest probabilities for predicting Spring SAT-10 were 33% and 35% for fall Nonsense Word Fluency and 37% and 36% for winter Nonsense Word Fluency in treatment and comparison conditions, respectively. Negative posttest probabilities were 13% and 16% for fall Nonsense Word Fluency and 10% and 12% for winter Nonsense Word Fluency in treatment and comparison conditions, respectively. Altogether, these statistics indicated that across concurrent and lag time administrations, instructional context did not appear to substantially alter the accuracy of Nonsense Word Fluency for ruling in or ruling out reading difficulties in individual students. Across contexts, in the current study Nonsense Word Fluency proved inadequate for ruling in reading difficulties on SAT-10 and was borderline adequate to inadequate for ruling out reading difficulties on SAT-10.

IV. DISCUSSION

As schools increasingly adopt early literacy CBMs as universal screeners within their MTSS-R frameworks, an argument-based approach to test validation is essential to determine how accurate these screeners are for (a) evaluating the overall effectiveness of a reading system (discriminative ability) and (b) predicting the likelihood of individual students having future reading difficulties (predictive ability). It is only through expressly evaluating a screener for its intended purpose(s) across instructional contexts that education researchers will be able to provide sufficiently nuanced recommendations to educators around when a screener is appropriate for making instructional decisions related to each of these purposes.

A crucial consideration in screener diagnostic accuracy within MTSS-R is the extent to which lag time between administrations of a screener and outcome measure may alter the diagnostic accuracy of early literacy CBMs and thus impact educators' interpretations of CBMs' use for both discriminative and predictive purposes. Medical research suggests that a lag between administrations of screeners and outcome measures can result in a "treatment paradox", in which individuals move from the population of individuals with the condition to the population of individuals without the condition due to the introduction of a "treatment" based on screening results, thus altering diagnostic accuracy estimates. The potential impact of the treatment paradox on screeners' diagnostic accuracy is especially important to study in the context of MTSS-R, where screeners are intentionally utilized to assign students to intervention based on screening results with the goal of improving these students' reading skills, so they no longer demonstrate reading risk prior to administration of an end-of-year outcome measure. It

would be expected that within the context of MTSS-R, the length of time between administrations of a reading screener and an end-of-year outcome measure would meaningfully alter the diagnostic accuracy of early literacy CBMs for predicting risk status on an outcome measure due to the strategic provision of supplemental instruction to students at risk.

Further, it would be expected that the effectiveness of the instruction that students receive would differentially alter a screener's test score interpretations and uses across lag times within an MTSS-R context. Specifically, poorer overall diagnostic accuracy statistics would be expected in contexts with stronger supplemental instructional effectiveness, particularly as there is greater lag time between screener and outcome measure administrations. The current study explicitly examined differences in an early literacy CBM's test score interpretations and uses across different lag times between administrations of a screener in contexts with varying instructional effectiveness.

The Impact of Lag Time on Overall Test Score Interpretations and Uses

The first major purpose of the current study was to evaluate the extent to which an early literacy CBM's test score interpretations and use for discriminative and predictive purposes varied based on the length of time between administrations of a screener on two different outcome measures—one proximal measure of oral reading fluency and one distal measure of overall reading achievement (SAT-10). As defined by Kane (2013)'s argument-based approach to test validation, a thorough test evaluation must include evidence for both the test's interpretations and uses.

Research Question 1 and 1a: Overall Test Scores Interpretations

DIBELS 6th Edition Nonsense Word Fluency was moderately to strongly associated with measures of reading fluency and overall reading achievement across time points, providing evidence for the screener's test score interpretations. Concurrent fall correlations were strongest across outcome measures, while concurrent spring correlations and fall and winter predictive correlations were similar and slightly weaker, though still moderate to strong, for Nonsense Word Fluency predicting both Oral Reading Fluency and SAT-10. These findings suggest that across students' first grade year, educators can confidently interpret student performance on these tests as being indicative of their skills in important reading areas.

One surprising finding in the current study was that for SAT-10, spring concurrent correlations were most similar to predictive correlations rather than fall concurrent correlations. Concurrent correlations are typically expected to be stronger than predictive correlations given that measures are administered at approximately the same timepoint. Yet the current study demonstrates that this may not always be the case. Concurrent spring correlations in the current study may have been weaker than expected due to the specific skills assessed on the DIBELS 6th Edition Nonsense Word Fluency measure. Follow-up distributional analyses indicate that there was a negative skew for spring Nonsense Word Fluency with many students obtaining scores of 120+ Correct Letter Sounds. Thus, at the end of first grade, there was a ceiling effect on DIBELS 6th Edition Nonsense Word Fluency performance such that the test may not have been effective at rank ordering student performance in a similar way to SAT-10. This finding was likely due to the fact that DIBELS 6th Edition Nonsense Word Fluency is comprised

of only vowel-consonant and consonant-vowel-consonant words, a decoding skill that at the end of first grade most students have already mastered. Thus, it may be important to utilize a decoding screener that includes higher-level decoding skills, such as long vowel sounds and r-controlled vowels, at the end of first grade. Updated versions of screeners such as DIBELS 8th Edition include these more advanced sound-spelling patterns.

Research Question 2 and 2a: Overall Discriminative Ability

Overall Appropriateness for Discriminative Purposes. In order to successfully evaluate the effectiveness of a current reading system within the context of MTSS-R, Nonsense Word Fluency must be able to accurately differentiate between students with and without reading difficulties, known as the test's discriminative ability. Nonsense Word Fluency demonstrated varying overall diagnostic accuracy for this purpose, with AUCs ranging from .80-.89 for Oral Reading Fluency (reasonable to very good overall accuracy) and .73-.81 for SAT-10 (poor to reasonable overall accuracy). Given an optimal cut score for risk that prioritized sensitivity of .90 or higher, Nonsense Word Fluency demonstrated below adequate specificity, with values ranging from .39-.59 for Oral Reading Fluency and .26-.39 for SAT-10.

These findings suggest that overall, when using an early decoding screener to evaluate the effectiveness of a school's reading system, it may not be feasible to expect that the screener will accurately classify most students with and without reading difficulties. Prioritizing a sensitivity rate of .90 or higher makes intuitive sense, as educators generally would prefer to inadvertently provide supplemental intervention to students who do not need it than to withhold intervention from students who are in need. However, in many school contexts sufficient resources are not available to provide strong

intervention supports to all students who are classified as at risk on a universal screener. In these cases, intervention supports are often diluted as schools increase the size of small groups, decrease the amount of time individual students spend in intervention groups, or allocate staff untrained in literacy instruction to teach intervention groups.

Educators may instead want to consider identifying a risk cut score that strikes a more equitable balance between sensitivity and specificity. For example, a cut score could be chosen that maximizes specificity given a sensitivity of .80 or higher, or by maximizing the combination of sensitivity and specificity, known as the Youden index (Smolkowski & Cummings, 2015). The optimal balance will depend on the instructional resources available within a school setting and should help to make systems-level instructional decisions that maximally serve students who may be at risk for reading difficulties while not overtaxing the system.

These findings should be considered in light of the fact that data were not available for students performing below the 10th percentile on fall SAT-10 in the current study. These students represent the population of individuals with the most intensive reading need, and it is expected that a majority of these students would remain below the 40th percentile on outcome measures at the end of the school year despite receiving effective reading intervention. The impact of the severity of a condition on diagnostic accuracy statistics has been widely documented in medical research, where it has been shown that in contexts with many individuals with very severe or very mild cases of a condition a screener will demonstrate stronger diagnostic accuracy (Leeflang et al., 2009). Applied to an educational context, it can be assumed that in a setting with many students with very low or very high skills, diagnostic accuracy will improve. Because

data were not available for the lowest performing students in the current sample, it would be expected that the overall diagnostic accuracy of Nonsense Word Fluency would be stronger in a typical school setting.

Variation in Discriminative Ability Based on Lag Time. At the same time, these findings must be considered through the lens of an argument-based approach to test validation. For the purpose of evaluating the overall effectiveness of a school's reading system, schools and districts rely on screeners to provide an accurate estimation of the *current* prevalence of reading difficulties in a given classroom, grade level, school, or district overall to support with instructional planning and to determine whether instructional supports are working as intended. Thus, a strategic evaluation of an early literacy screener for this purpose will consider the discriminative ability of a fall, winter, and spring screener for predicting risk status on a *concurrently administered* outcome measure as the most appropriate evidence for this test score use.

Findings from the current study demonstrated that lag time may in fact alter diagnostic accuracy statistics associated with a screener's discriminative ability (i.e., AUC, sensitivity, and specificity) for predicting reading difficulties on a measure of oral reading fluency and an overall measure of reading achievement. Specifically, as lag time increased, Nonsense Word Fluency became increasingly poor at accurately identifying students who did not have reading difficulties. In other words, with increased time between administrations of Nonsense Word Fluency and each outcome measure, specificity rates decreased, with meaningful differences in specificity between fall, winter, and spring administrations of Nonsense Word Fluency for predicting spring Oral

Reading Fluency and between fall and spring administrations of Nonsense Word Fluency for predicting spring SAT-10.

Findings from the current study suggest that if educators' intention is to use a screener for evaluating the current health of their reading system, using diagnostic accuracy statistics derived from lag time administrations may inaccurately represent the screener's appropriateness for this purpose. In the field of education, sensitivity and specificity values have been assumed to be reasonably static, and thus generalizable, across lag times (Kilgus et al., 2014). The current study demonstrates that this assumption may not be true in all instances.

Future Research Directions. Researchers studying literacy screeners have often chosen to evaluate the diagnostic accuracy of a screener administered in the fall for discriminating between students with and without reading difficulties at the end of the school year. In fact, until recently prominent organizations such as the National Center on Intensifying Interventions have required that screener evaluations demonstrate a lag time of at least three months between administrations of the screener and outcome measures in order to be considered for review (NCII, 2018). However, given that in general the purpose of using an early literacy screener for discriminative purposes is to identify proportions of students who *currently* do and do not have reading difficulties, it may not be necessary or even prudent to design diagnostic accuracy studies with lag time between screener and outcome measure administrations. Future research is needed to compare diagnostic accuracy statistics for other commonly used early literacy screening measures when administered at varying lag times to determine whether this finding can be generalized across screening tools. In the meantime, test developers should explicitly

report any lag time that occurred between screener and outcome measure administrations along with screener diagnostic accuracy statistics and recommended cut-scores for risk.

Implications for Educators. When examining the current effectiveness of their reading system, educators are likely most interested in understanding a screener's ability to accurately discriminate between students with and without current reading difficulties so that decisions can be made about the effectiveness of the instruction and intervention practices that are being utilized and how to effectively allocate limited school resources. For this purpose, examining concurrent fall, winter or spring sensitivity and specificity values is likely most relevant for choosing an appropriate screening tool. The current study indicates that it may be reasonable to adopt a screener that targets decoding skills such as Nonsense Word Fluency in first grade for making low-stakes systems-level decisions, particularly if an equitable balance between sensitivity and specificity values is prioritized when choosing a cut score for risk.

Research Question 3 and 3a: Overall Predictive Ability

Overall Appropriateness for Predictive Purposes. To appropriately assign students to supplemental and intensive intervention within MTSS-R, Nonsense Word Fluency must also be able to accurately predict an individual student's likelihood of having reading difficulties, described as the test's predictive ability. Nonsense Word Fluency demonstrated poor to reasonable predictive ability for determining a student's likelihood of having reading difficulties given an "at risk" screening result, with positive likelihood ratios ranging from 1.48 to 2.20 for Oral Reading Fluency and 1.23 to 1.48 for SAT-10. These values corresponded to positive posttest probabilities ranging from 18% to 25% for Oral Reading Fluency and 35% to 42% for SAT-10. These findings suggested

that given base rates of reading difficulties ranging from 12% to 13% on Oral Reading Fluency and 28% to 33% on SAT-10, Nonsense Word Fluency generally was not adequate for ruling in reading difficulties.

Nonsense Word Fluency demonstrated slightly stronger predictive ability for predicting a student's likelihood of reading difficulties given a "not at risk" screening result, with negative likelihood ratios ranging from .17 to .26 for Oral Reading Fluency and .26 to .35 for SAT-10. These statistics corresponded to negative posttest probabilities ranging from 2% to 4% for Oral Reading Fluency and 10% to 13% for SAT-10, indicating that given the relatively low base rates of reading difficulties across outcome measures, Nonsense Word Fluency was generally appropriate on Oral Reading Fluency and borderline adequate on SAT-10 for ruling out reading difficulties.

Within MTSS-R, screeners are not intended to make high stakes decisions such as diagnosing students with reading disabilities, but rather lower stakes decisions such as identifying those students who would most benefit from supplemental reading supports. Thus overall, findings from the current study support previous evidence suggesting that early literacy screeners administered within MTSS-R may not be sufficiently accurate for ruling in reading difficulties (e.g., VanDerHeyden, 2018), but may provide valuable information to help rule out reading difficulties for students in 1st grade, particularly on a reading fluency outcome measure.

It is also important to note that in the current study, reading difficulties were defined by performance at or below the 40th percentile on each outcome measure. Thus, it is likely that Nonsense Word Fluency demonstrated a stronger ability to rule out than rule in reading difficulties in the current study because students had to demonstrate fairly

strong reading skills to be classified as "not at risk" (i.e., perform as well as or better than 40% of students nationwide). It would be expected that had reading difficulties been re-defined as performance at or below the 10th or 20th percentile, which the National Center on Intensive Intervention identifies as the appropriate criteria for identifying students in need of intensive intervention, Nonsense Word Fluency would demonstrate stronger rule in ability (NCII, 2020). Further, given that the lowest performing students were not included in current study analyses, it would be expected that Nonsense Word Fluency would do a better job of ruling in reading difficulties within a typical school setting which is representative of students with intensive reading needs. Future research should examine Nonsense Word Fluency's predictive ability across multiple risk criteria and student populations to determine the tool's appropriateness for both ruling out and in reading difficulties.

Variation in Predictive Ability Based on Lag Time. Similar to findings regarding Nonsense Word Fluency's discriminative ability, the extent to which Nonsense Word Fluency was appropriate for predictive purposes varied based on lag time between test administrations. Thus, study results suggest that it is vital for researchers and practitioners alike to use an argument-based approach to test validation to comprehensively consider a screener's appropriateness for their intended purposes within MTSS-R. Across outcome measures, positive likelihood ratios in particular grew meaningfully poorer as lag time increased in the current study, suggesting that with greater lag time between test administrations, Nonsense Word Fluency screening results grew less helpful for ruling in reading difficulties.

That increased lag time led to poorer predictive ability has important implications, given that educators typically rely on screeners within MTSS-R to accurately predict the likelihood of individual students developing reading difficulties *in the future*. For this purpose, fall and winter screenings must provide a reasonably accurate prediction regarding whether or not a student will perform poorly on an end-of-year outcome measure. This helps educators assign supplemental intervention to all students who would end up performing below grade-level expectations without it and to withhold intervention from students who do not need it to meet end-of year grade level expectations. Findings from the current study suggest that it may not be appropriate to rely alone on an early literacy screener administered in the fall or winter to provide a highly accurate likelihood of end-of-year reading difficulties for an individual student within the context of MTSS-R, as in many cases the screener will provide an inaccurate prediction.

At the same time, meaningful differences in likelihood ratios across lag times did not necessarily result in major differences in the probability of an individual student having current or future reading difficulties within the current study context. For instance, in the current study positive posttest probabilities only varied by up to 7% for both Oral Reading Fluency (range = 18% to 25%) and SAT-10 (range = 35% to 42%), while negative posttest probabilities were even less variable, varying by no more than 2% (range = 2% to 4%) for Oral Reading Fluency and 3% (range = 10% to 13%) for SAT-10 across time points and lag times. Posttest probabilities did not vary drastically enough to warrant different instructional decisions (e.g., provide intervention, provide follow-up testing, or withhold intervention) based on probabilities obtained at specific time points or lag times, as recommended by VanDerHeyden (2013). Thus, given the relatively low

base rates of reading difficulties in the current study, meaningful differences in likelihood ratios across lag times were not necessarily large enough to recommend any changes to how educators use or interpret posttest probabilities based on time of year.

It was hypothesized that decreasing base rates of reading difficulties from beginning to end of year would result in Nonsense Word Fluency becoming a progressively stronger tool for ruling out reading difficulties and a poorer tool for ruling in reading difficulties. Instead, neither positive nor negative posttest probabilities substantially changed across the year in the current study. This may have been due to the fact that base rates of reading difficulties did not vary greatly across the school year as defined by either outcome measure. Specifically, base rates moved from 12% in the winter to 13% in the spring on Oral Reading Fluency, and from 33% in the fall to 28% in the spring on SAT-10. This finding demonstrates that in the context of a randomized controlled trial where students who were identified as at risk for reading difficulties were assigned to daily Tier 2 intervention, base rates of reading difficulties as defined by highly-regarded outcome measures may not change dramatically. Thus, base rates may be more static across the school year than had been predicted.

It should be noted that because data were not available for any students who performed below the 10th percentile on fall SAT-10, base rates of reading difficulties were not high at any time point in the current study. Study findings must be considered in light of this limitation. It is possible that had these most at risk readers been included in study analyses, base rates of reading difficulties and thus post-test probabilities would have varied more. However, given that students with intensive reading needs are typically less likely to make substantial reading progress across a school year (Toste et al., 2014),

it is likely that though their inclusion would have resulted in an overall increase in base rate of reading difficulties, it would likely not have resulted in any substantial change in base rate across the year.

At the same time, it would be expected that in contexts with higher beginning of year base rates of reading difficulties, particularly in contexts with many students who are reading slightly to moderately behind their grade level peers, provision of supplemental intervention to all at risk students would result in greater base rate shifts across the school year. In these cases, posttest probabilities may be more profoundly altered, particularly for predicting an individual student's likelihood of *current* reading difficulties. Future research should consider the extent to which base rates of reading difficulties shift across the year in typical school settings, and whether posttest probabilities vary more drastically in contexts with widely shifting base rates.

Future Research Directions. The current study demonstrates that particularly in settings with low base rates of reading difficulties, an early literacy screener such as Nonsense Word Fluency may not provide enough information to accurately predict an individual student's likelihood of having current or future reading difficulties. Previous studies with similar findings have thus concluded that schools with reasonably low base rates of reading difficulties may be best off eliminating the use of screeners altogether for this predictive purpose, instead using prior year end-of-year statewide assessments to identify students in need of supplemental intervention (e.g., VanDerHeyden et al, 2018). However, these studies have focused on students in later elementary school when statewide assessments are typically mandated, with the argument that the additional time spent on an assessment that adds little meaningful information may be counterproductive

to students' reading outcomes. Most students in kindergarten through 2nd grade generally are not required to take a statewide assessment, however, so this recommendation may be less meaningful to early elementary educators.

Instead, findings from the current study suggest an urgent need for researchers to identify screening approaches that improve on the predictive ability of early literacy screeners in kindergarten through second grade. Most existing studies have focused on improving screeners' diagnostic accuracy via a "gated" screening approach, where the added benefit of additional reading-related screening or progress monitoring data is evaluated following an initial "at risk" screen (Catts et al., 2015; Compton et al, 2012; Gilbert et al., 2012). Studies that have examined a gated screener's impact on predictive ability have found some promise in the approach, with gated screening resulting in statistically significant decreases in false positive screening results (Van Norman et al., 2017). However, these studies have all targeted upper elementary students and more research is needed on the impact of gated screening approaches on the predictive ability of screeners in early elementary.

Researchers have also come to recognize in recent years that reading disabilities such as dyslexia are caused by a complex host of variables, such that it is difficult to attribute an individual's reading difficulties to any one particular deficit (Catts & Petscher, 2020). Emerging research further indicates that adding highly correlated measures to a screening battery may result in smaller reductions in a screener's false positive rate than adding measures that are less correlated (VanNorman et al., 2018). Future studies could examine whether the addition of non-reading related screening measures associated with reading disabilities, such as a student's mindset, behavior, or

family history of reading disabilities, may improve screeners' predictive ability for the purpose of identifying students' future likelihood of reading difficulties within MTSS-R (Catts & Petscher, 2020; Greulich et al., 2014). Studies could also consider whether calculating posttest probabilities using interval likelihood ratios, which partition screening scores into more than two categories of risk, adds meaningful information for instructional decision-making for certain groups of students as suggested by emerging research (Klingbeil et al., 2019). The current study demonstrates that these more nuanced approaches to classifying students may be especially important in contexts such as the current study, where students demonstrate moderate, but not substantial risk for reading difficulties. Research suggests that screeners may be poorest at classifying this group of students (Johnson et al., 2009).

Implications for Educators. Within MTSS-R, educators typically use screening scores to assign students to supplemental reading supports based on their risk for future reading difficulties. The current study suggests that in a context with a low base rate of reading difficulties, a single early literacy screener targeting students' decoding skills in first grade may not be sufficient to provide a highly accurate prediction of an individual student's likelihood of developing future reading difficulties. For example, in the current study, the rate of reading difficulties for students classified as "at-risk" on the screener ranged from 17 to 42%, indicating that across lag times the screener inaccurately predicted that a student would have reading difficulties over half of the time.

It is also important for educators and parents to recognize that though a screener may classify a student as at-risk and in need of supplemental reading supports, an "at risk" screening result does not necessarily indicate that a student is going to have ongoing

reading difficulties. Similarly, students who are classified as "at risk" are not all equally likely to end up with reading difficulties. This is particularly important to consider in the context of the widespread adoption of state dyslexia screening legislation that requires schools to screen students for dyslexia risk and to notify parents whose children are classified as "at risk" on the dyslexia screener (National Center on Improving Literacy, 2020).

Educators are responsible for using screening data to both make instructional decisions and share individual students' screening results with parents. Thus, it is critical that educators know how to accurately interpret screening scores such that the interpretations and decisions made based on these scores are justifiable. In other words, actions taken based on screening scores must demonstrate strong consequential validity. The current study suggests that while an early literacy CBM provides some helpful information for making low-stakes decisions such as assigning a student to supplemental intervention, a screening score alone should not be used for higher-stakes decision making.

Educators should make use of this knowledge as they use screeners to support their decision-making processes. For instance, when sharing information about a student's dyslexia risk with parents, it may behoove educators to provide a probability of the child's likelihood of having dyslexia currently versus in the future based on screening results, given the historical base rate of reading difficulties at the school. Additionally, it may be helpful to consider the potential consequences of providing students with supplemental or intensive intervention when deciding on necessary supports for students classified as "at risk". For example, educators should discuss the extent to which

supplemental or intensive intervention will limit students' access to core instruction and how frequently student data will be utilized to move students between tiers of instruction across the school year. In contexts where strategic or intensive intervention supplants core instruction or where students are infrequently moved between tiers of instruction, instructional decisions may be considered higher-stakes and additional sources of data beyond screening results should be utilized.

To support with the triangulation of multiple data sources in conjunction with screening scores for making instructional decisions about individual students, some researchers have suggested the adoption of nomograms (Pendergast et al., 2018). A nomogram is a practitioner-friendly tool frequently used in medicine and mental health for calculating individuals' likelihood of having a condition given the known population base rate of the condition in individuals with similar characteristics such as race/ethnicity and age. In education, a nomogram could allow for more nuanced calculations of a student's pretest probability of reading difficulties based on the integration of multiple sources of data such as teacher ratings, lesson mastery or progress data, and family history of reading difficulties in addition to early literacy screener scores.

Instructional Effectiveness and Test Score Interpretations and Uses

The second major purpose of the current study was to determine the extent to which the effectiveness of instruction being provided to students identified as "at risk" on an early literacy screener differentially altered the impact of lag time on the screener's test score interpretations (i.e., concurrent and predictive validity) or uses for discriminative (i.e., AUCs, sensitivity and specificity) and predictive (i.e., likelihood ratios and posttest probabilities) purposes. In the current study, the impact of instructional

effectiveness on test score interpretations and uses was isolated by evaluating an early literacy screener in the context of a randomized controlled trial where schools were randomly assigned to receive either a multi-tiered reading intervention found to be effective for improving the reading skills of students identified as at risk for reading difficulties, or business-as-usual core and supplemental reading instruction. Thus, it can be assumed that any differences in test score interpretations and uses between conditions in the current study were due to the increased effectiveness of the intervention being provided rather than some other third variable. If instructional effectiveness differentially impacts the degree to which a screener's test score interpretations and uses are affected by lag time, researchers and educators may need to consider not only lag time between test administrations, but also the specific instructional supports they are providing when using an argument-based approach to evaluating the accuracy of a screening tool for their intended purposes within the context of MTSS-R.

Overall Test Score Interpretations and Uses

The impact of instructional effectiveness on Nonsense Word Fluency's test score interpretations or uses for discriminative or predictive purposes when predicting either Oral Reading Fluency and SAT-10 risk status varied. In most cases, there were no meaningful differences in AUCs, sensitivity or specificity values, or likelihood ratios. This finding was not altogether unexpected given that in the current study, overall instructional effectiveness did not vary substantially between conditions when taking Tier 1 and 2 students into account. That is, though at-risk students in the treatment condition consistently outperformed their at-risk peers in the comparison condition, effect sizes were somewhat small in many cases (g = 0.12 for SAT-10 Total Reading; g = 0.25 for

ORF) and not significant for SAT-10 (Fien et al., 2020), and mean reading scores did not substantially vary between treatment and comparison conditions for the overall sample of Tier 1 and 2 students. Thus, it is possible that meaningful differences in correlations and diagnostic accuracy were not observed between treatment and comparison conditions in the current study because the effectiveness of instruction students received was not different enough to result in substantial changes to these statistics. In both conditions, all students received an average of 90 minutes of Tier I instruction using a core reading program, while at-risk students received an additional 30 minutes of daily Tier II intervention; thus, the amount of reading instruction all students received was aligned with recommended MTSS-R practices (Gersten et al., 2009).

Follow-up descriptive analyses of the data affirm this hypothesis: for the sample of Tier 1 and 2 students in the current study, the percent of students who changed risk status from beginning to end of year (e.g., moved from the population of students with "reading difficulties" to "no reading difficulties" or vice versa) was not substantially different across conditions. For example, 66.0% versus 59.8% of students changed from "reading difficulties" to "no reading difficulties" status, while 11.4% versus 7.5% of students changed from "no reading difficulties" to "reading difficulties" status on Oral Reading Fluency from winter to spring for treatment and comparison conditions, respectively. Similarly, 45.3% versus 46.3% of students changed from "reading difficulties" to "no reading difficulties" status, while 24.7% versus 26.7% of students changed from "no reading difficulties" to "reading difficulties" status from fall to spring for treatment and comparison conditions, respectively on SAT-10.

Future Research Directions. In practice, MTSS-R is implemented with a great degree of variability across schools and districts (Berkeley et al., 2020; Gersten, Jayanti, et al., 2017). Larger differences in the percentage of students who change reading status across the year would be expected among MTSS-R settings with more widely varying instructional effectiveness. Thus, the difference in instructional effectiveness between the two conditions in the current study may not be sufficiently representative of the actual variance in instructional effectiveness demonstrated by schools outside of the context of a research study and findings from the current study may not necessarily generalize across all instructional contexts.

Future research could use simulation methodology to systematically examine how discriminative and predictive diagnostic accuracy vary across contexts with different degrees of instructional effectiveness. Because lag time is expected to impact diagnostic accuracy most in contexts with more drastic changes in student rank order across administrations of screeners and outcome measures, it would be predicted that diagnostic accuracy would look worst in settings in which core instruction is relatively ineffective and supplemental intervention is relatively effective. Simulation studies could systematically test this hypothesis by evaluating a known screening assessment across datasets that mimic these different contexts. Should this hypothesis be borne out, researchers may need to consider the impact of instructional context when designing screener evaluation studies and report on this contextual information when presenting diagnostic accuracy findings.

It is also important to note that there are multiple ways of defining "instructional effectiveness" and that varying definitions may result in different degrees of impact on a

screener's diagnostic accuracy. In the current study, high instructional effectiveness was defined by enrollment in the treatment condition, in which an intervention was provided that on average improved reading outcomes for at-risk students above and beyond business-as-usual Tier II interventions. However instructional effectiveness varied somewhat from classroom to classroom within treatment and comparison conditions. Quality of explicit instruction data collected in all classrooms in the original ECRI study indicated that though the mean quality of explicit instruction score was significantly higher in the treatment condition (0.89) than in the comparison condition (0.49), quality of explicit instruction scores varied within each condition, with standard deviations of 0.17 and 0.25 in the treatment and comparison conditions, respectively.

Future research could examine the extent to which a screener's diagnostic accuracy varies by classroom-level instructional effectiveness. For instance, diagnostic accuracy statistics could be compared for classrooms demonstrating high versus moderate versus low instructional effectiveness as indicated by implementation fidelity data across the school year, regardless of experimental condition. Meaningful differences in screener diagnostic accuracy across classrooms may suggest the need for researchers and educators alike to place greater import on the impact that teacher instruction has on a student's likelihood of future reading difficulties.

Meaningful Differences Between Conditions

In the current study, meaningful differences were observed in optimal cut-scores for risk for several timepoints and lag times. Specifically, cut-scores for winter and spring Nonsense Word Fluency predicting winter and spring Oral Reading Fluency varied between conditions by 3 to 10 correct letter sounds, while cut-scores for winter and

spring Nonsense Word Fluency predicting spring SAT-10 varied between conditions by 8 to 14 correct letter sounds. In other words, across these time points and lag times, students in the treatment condition needed to demonstrate stronger performance on Nonsense Word Fluency in order to be classified as not at-risk on Oral Reading Fluency or SAT-10 than students in the comparison condition.

Additionally, the optimal cut point for spring Nonsense Word Fluency predicting both spring Oral Reading Fluency and spring SAT-10 resulted in meaningfully stronger specificity values and likelihood ratios in the treatment condition than the comparison condition, though no meaningful differences in AUC values were observed. This indicated that Nonsense Word Fluency appeared to be a stronger tool for both discriminative and predictive purposes in a context with higher instructional effectiveness when predicting concurrent spring risk status. Further, when an optimal cut score was chosen that prioritized maintaining sensitivity at or above .90, cut scores varied by 10 or more points across conditions, with optimal cut scores of 61.50 and 71.50 for Nonsense Word Fluency predicting Oral Reading Fluency in treatment and comparison conditions, respectively, and optimal cut scores of 60.50 and 74.50 for Nonsense Word Fluency predicting SAT-10 in treatment and comparison conditions, respectively.

These findings ran counter to two of the current study's hypotheses: (1) that diagnostic accuracy differences between conditions would be greater as lag time increased, and (2) that when diagnostic accuracy statistics were meaningfully different between conditions, they would appear stronger in the comparison condition. One explanation for these counterintuitive findings is that the supplemental intervention provided to students in the high instructional effectiveness (ECRI treatment) condition

may have placed more of an emphasis on teaching students how to generalize decoding skills to reading fluency and reading comprehension skills. For example, within a typical Tier 2 supplemental ECRI lesson, students not only receive instruction on sound-spelling patterns and word blending, but also gain practice with reading connected decodable texts. Thus, students in the current study who responded well to the ECRI Tier 2 intervention may have simultaneously built skills across foundational areas of decoding, reading fluency, and comprehension, resulting in similar rank ordering of student skills on both the screener and both outcome measures.

In contrast, it is possible that students in the comparison condition received instruction that was not as strategic about integrating foundational skills, and so improvements in decoding skills may not have led to substantial improvements in reading fluency and comprehension. If this were the case, Nonsense Word Fluency would do a poorer job of classifying students as "not at risk" in the comparison condition because in this condition, Nonsense Word Fluency scores would be less closely correlated with Oral Reading Fluency and SAT-10 scores for students who had received Tier 2 intervention. Information on the specific content taught within Tier 2 intervention in the comparison condition was not available, however teachers reported that they used a variety of published, standardized protocol intervention materials and teacher-developed materials to teach these groups.

At the same time, despite differences in cut-scores and specificity rates it is important to note that ROC curves were visually similar and overall AUCs (as well as sensitivity and specificity rates in most cases) were nearly identical for all combinations of screeners and outcome measures across conditions. Thus, it is possible that these

varying cut-scores and non-overlapping specificity values may not actually have been the result of meaningful differences in screener accuracy between conditions, but rather the result of a jagged ROC curve due to the small sample size which made specificity values look particularly different at the cut score that corresponded to a sensitivity value of .90 or higher. As ROC analyses are conducted on progressively larger sample sizes, ROC curves grow increasingly smooth, and less jagged. Though the sample size for each ROC analysis in the current study was between 700 and 800, it was below the ideal sample size for a test with specificity rates as low as Nonsense Word Fluency demonstrated in the current study for producing highly precise estimates (Malhotra & Indrayan, 2010), and so sensitivity and specificity estimates for specific cut scores were less precise than would be ideal.

Researchers and practitioners should be sensitive to the fact that differences in sensitivity and specificity rates can occur for measures with similar overall accuracy, particularly when ROC analyses are conducted on smaller samples. This may result in sensitivity and specificity values that seem less than ideal for certain risk cut scores. It is important to look at the entirety of a ROC curve when choosing an ideal cut score based on an optimal balance of sensitivity and specificity for a specific screener in a given setting (Smolkowski & Cummings, 2015). More diagnostic accuracy research is needed to determine whether this finding is replicable in other instructional contexts and with other early literacy screeners.

This finding also has implications for how researchers and educators interpret screeners' diagnostic accuracy given a larger phenomenon which has appeared in recent decades across reading intervention studies —a general improvement in standard

supplemental reading instructional practices which has resulted in novel supplemental reading interventions demonstrating smaller overall effect sizes across time (Bakker et al., 2019). As the quality of supplemental reading instruction improves nationwide, it is particularly important to consider the instructional environment in which optimal cut scores were originally derived for a screening measure when deciding on an appropriate screening tool, and how much that context differs from current instructional practices.

Implications for Educators. Findings from the current study suggest that small differences in instructional effectiveness may meaningfully impact an early literacy screener's test score interpretations and uses for discriminative or predictive purposes, most notably the optimal cut-scores for risk. Thus, educators should take caution in relying on established cut-scores and diagnostic accuracy statistics of published screeners when adopting a tool for their school setting. For example, in the current study cut-scores for risk varied substantially from the original DIBELS 6th Edition benchmark cut-scores: in the fall, the established DIBELS 6th cut-score was 25 correct letter sounds, or 3.5 to 7.5 fewer correct letter sounds than the optimal cut-score for treatment and comparison conditions in the current study when predicting to either outcome measure. In the winter, the established cut-score was 54, while in the spring it was 71 correct letter sounds. At both of these timepoints, the established DIBELS 6th Edition cut-scores also varied from the optimal cut-scores for risk identified across conditions and outcome measures, from as little as a decrease of .5 correct letter sounds for spring Nonsense Word Fluency predicting spring Oral Reading Fluency to as much as an increase of 10.5 correct letter sounds for spring Nonsense Word Fluency predicting spring SAT-10. Clearly, it may be prudent to carefully consider the context in which established cut-scores are derived,

including the normative sample, the instructional setting, and the chosen definition of reading risk, and to identify screener cut-scores that most closely align with their own instructional context and purposes for instructional decision making.

At the same time, despite large differences in optimal cut-scores for risk and some meaningful differences in specificity between conditions, post-test probabilities remained fairly similar across conditions, likely due in part to similar base rates of reading difficulties across instructional contexts in most cases. This finding suggests that for the purpose of predicting an individual student's likelihood of reading difficulties, differences in a screener's sensitivity, specificity, and optimal cut score may not meaningfully alter post-test probabilities as much as the existing base rate of reading difficulties in their context. In other words, educators who are using a screener to predict an individual student's likelihood of developing reading difficulties may find that information about the proportion of students in their school who have historically had reading difficulties may more meaningfully change an individual's risk prediction that differences in sensitivity, specificity, or cut scores.

In fact, this phenomenon was observed in the current study. For example, despite meaningful differences in specificity and likelihood ratios across certain time points, lag times, and conditions for Nonsense Word Fluency predicting Oral Reading Fluency performance, positive post-test probabilities never varied by more than 8% (range = 17% to 25%), and negative post-test probabilities never varied by more than 2% (range = 2% to 4%) for the overall sample, where base rates of reading difficulties remained between 12 and 13%. In all cases, differences in post-test probabilities never warranted changes in decision making based on VanDerHeyden (2013)'s recommendations.

In contrast, when comparing diagnostic accuracy of Nonsense Word Fluency predicting Oral Reading Fluency performance versus SAT-10 performance in cases where specificity values and likelihood ratios were comparable but base rate varied widely, much larger differences in posttest probabilities were observed. For example, fall Nonsense Word Fluency predicting winter Oral Reading Fluency resulted in a specificity value of .41, 95% CI [.38, .44] and positive and negative likelihood ratios of 1.53, 95% CI [1.43, 1.63] and 0.24, 95% CI [0.16, 0.38], respectively, while fall Nonsense Word Fluency predicting fall SAT-10 resulted in a specificity value of .39, 95% CI [.36, .42] and positive and negative likelihood ratios of 1.48, 95% CI [1.39, 1.56] and 0.26, 95% CI [0.19, 0.34], respectively. Despite similar population-based statistics, positive and negative post-test probabilities were substantially different, with positive post-test probabilities of 17% and 42% and negative post-test probabilities of 3% and 11% for Oral Reading Fluency and SAT-10, respectively. These values corresponded to varying base rates of reading difficulties of 12% for Oral Reading Fluency and 33% for SAT-10.

In other words, varying base rates of reading difficulties contributed to vastly different likelihoods of student reading difficulty which would result in different instructional decisions based on VanDerHeyden (2013)'s recommendations. It can be inferred that using an early literacy screener with reasonable overall accuracy in a setting with low base rates of reading difficulties, as with the example of Oral Reading Fluency, will result in "at risk" screenings providing little useful information, whereas "not at risk" screenings will be quite useful for ruling out reading difficulties and identifying students who will become proficient readers without supplemental intervention. In contrast, the same screener used in a setting with a larger base rate of reading difficulties, as in the

case with SAT-10, will result in “at risk” screenings that provide more useful information for ruling in reading difficulties and identifying students who most likely need supplemental intervention to prevent reading difficulties, while a “not at risk” screening may not alone provide useful information for decision-making.

These findings indicate that educators should consider historical base rates of reading difficulties when determining how best to utilize early literacy screeners for making decisions about individual students in their setting. For instance, when using a screener in a setting with historically low base rates of reading difficulties, educators may need to prioritize allocating resources to collecting additional information on students who are classified as “at risk”. Depending on the number of students who fall into this category, this may include assigning students to short-term supplemental intervention, follow-up assessment, or simply monitoring student progress over time.

In contrast, when using an early literacy screener in a setting with historically high base rates of reading difficulties, it is likely that many students will need evidence-based supplemental reading supports to ensure end-of-year reading proficiency, including some students who have been classified as “not at risk”. In this case, it will be crucial that educators continue to closely monitor and provide high quality instructional supports even to students who were classified as “not at risk”. Thus, a focus on high quality and differentiated core instruction will be essential. For example, in addition to providing supplemental or intensive intervention to those students who have been classified as “at risk”, educators in this context should prioritize the use of explicit and systematic instruction in conjunction with the collection and use of in-program mastery data at Tier 1. This will enable them to closely monitor the progress of and provide differentiated

supports to students who were classified as “not at risk” but who may still be in danger of performing below end-of-year grade-level expectations.

These findings should be interpreted cautiously given that the degree of instructional effectiveness in the current study is likely not representative of many school-based contexts. In general, it would be beneficial for educators to consider contextual factors such as the effectiveness of instruction in their setting as well as historical base rates of reading difficulties when determining how best to utilize early literacy screeners for instructional decision making.

Study Limitations

Findings from the current study should be viewed in light of several limitations. First, because analyses were conducted using an existing dataset, it was not possible to conduct an examination of diagnostic accuracy statistics for certain timepoints. For example, an analysis of beginning of year Oral Reading Fluency data may have allowed for a better understanding of how shifting base rates of reading difficulties alter screener diagnostic accuracy across the year given that Oral Reading Fluency was more proximal to the reading instruction being provided. However, these data were not collected in the original ECRI study.

Similarly, the subtests that comprised the SAT-10 Total Reading score differed from the beginning to end of year, and so differences in diagnostic accuracy on SAT-10 may have partially been attributed to differences in the specific reading-related skills that were assessed at each time point. However, the subtests that contributed to the Total Reading score at each timepoint were specifically developed to be grade-appropriate and aligned with state and national standards (Pearson Education, 2018). Thus, though

different subtests were used at each time point, it can be assumed that fall and spring Total Reading scores were both intended to provide an accurate measure of overall reading achievement at their respective timepoints.

Further, in the original ECRI study, students who fell below the 10th percentile were excluded from the analytic sample; thus, screening and outcome data were unavailable for those students most at risk for reading difficulties. This restricted sample likely made Nonsense Word Fluency appear less accurate for both discriminative and predictive purposes than would be the case had these lower performing students been included in current study analyses. Thus, findings from the current study should be interpreted with this limitation in mind. Nevertheless, differences found in the current study across lag times and between instructional conditions would be expected to hold even with the addition of highly at-risk students, given the high threshold chosen for defining reading risk (i.e., below the 40th percentile).

Second, overall instructional effectiveness did not vary drastically across conditions in the current study for the overall sample including Tier 1 and 2 students, and as such meaningful differences in diagnostic accuracy statistics were not as large as expected. Larger differences in both a screener's discriminative and predictive ability would be expected in contexts where instructional effectiveness varies more greatly, as is commonly the case in school settings. Future studies could systematically vary instructional effectiveness at both Tier 1 and 2 to determine whether diagnostic accuracy statistics meaningfully change with increased lag time between administrations of screener and outcome measures. However, findings from the current study indicate that even small differences in overall instructional effectiveness may substantially alter

optimal cut-scores for risk and in some cases the discriminative and predictive ability of an early literacy screener.

Third, as previously mentioned, though sample sizes for each ROC analysis in the current study were larger than many existing diagnostic accuracy studies (e.g., Hintze et al., 2003; Nelson, 2008), a larger sample size was needed to estimate overall diagnostic accuracy with absolute precision, and in some cases confidence bounds around estimates were quite large. Future research should replicate this study's approach with a sample size in the thousands—for example, in their evaluation of the DIBELS 6th Edition measures, Smolkowski and Cummings (2015) were able to estimate confidence intervals of $\pm$.02 around decision thresholds and $\pm$.01 for AUC values with a sample size of approximately 4,000 students.

Conclusion

The current study demonstrates the importance of using an argument-based approach to evaluating early literacy screeners' test score interpretations and uses for both discriminative and predictive purposes. Specifically, researchers and educators alike should closely consider their screening purpose(s) when evaluating a screener for use within their MTSS-R setting. The current study indicates that educators are likely on solid ground when using an early literacy screener to evaluate the current reading skills of 1st grade students overall, particularly for the purpose of determining whether core instruction needs to target reading fluency, and by proxy, basic comprehension skills. However, the current study also indicates that educators should take caution when using an early literacy screener to predict whether an individual student will have reading difficulties, as predictions will likely frequently result in incorrect decisions. Educators

should use early literacy screening data in combination with other data sources whenever making high stakes decisions about individual students. Finally, the current study suggests that the effectiveness of instruction students receive across the school year may substantially impact optimal cut-scores for risk and in some cases diagnostic accuracy statistics. However, more research is needed to evaluate this finding across contexts with more widely varying instructional effectiveness.

APPENDIX

CORRELATIONAL ANALYSIS ASSUMPTIONS

Figure 14

Scatterplots of Standardized Predicted Values of Outcomes Regressed on Screening Measures

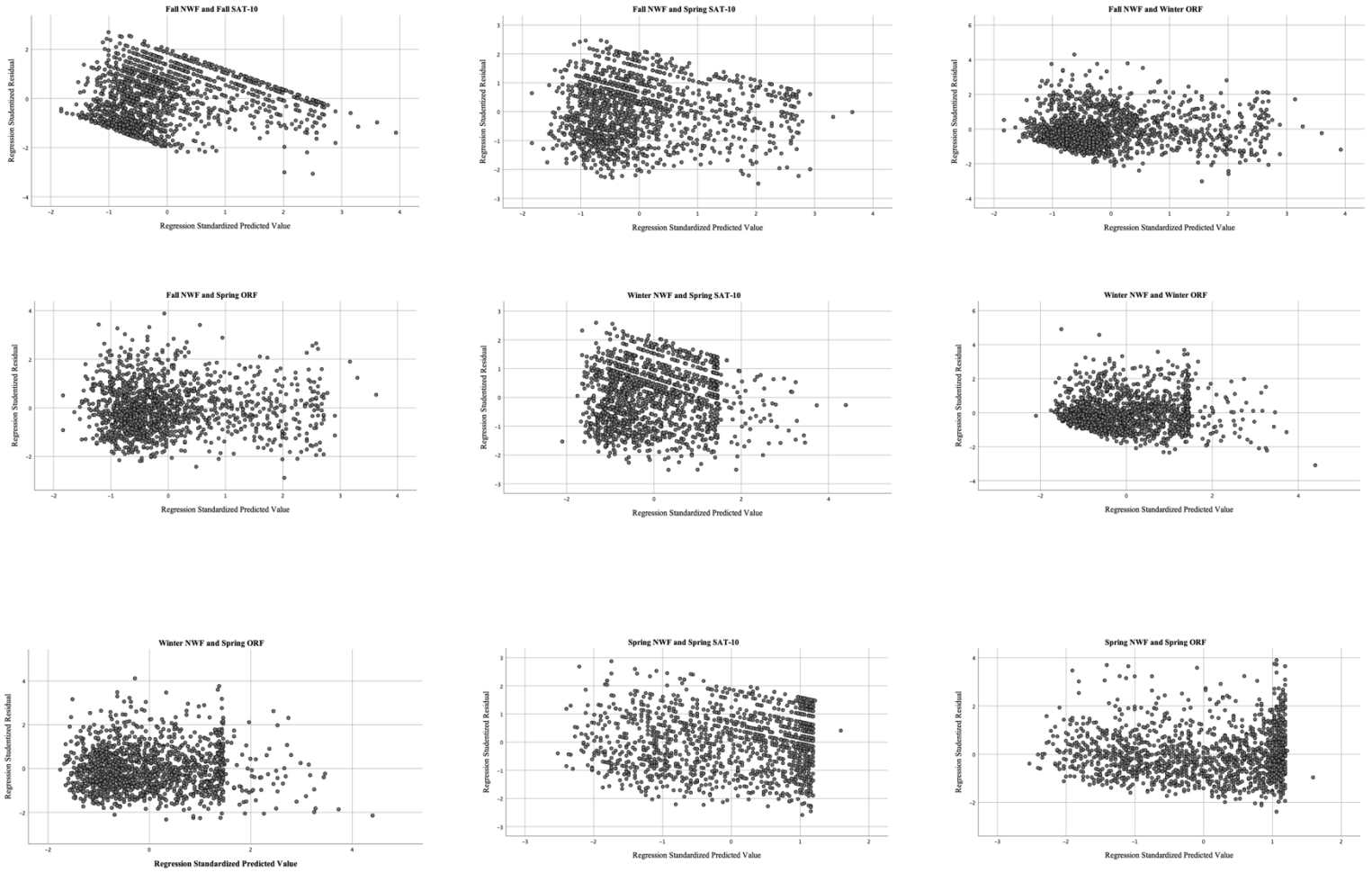


Figure 15

Histograms of Standardized Residuals for Screening and Outcome Measures

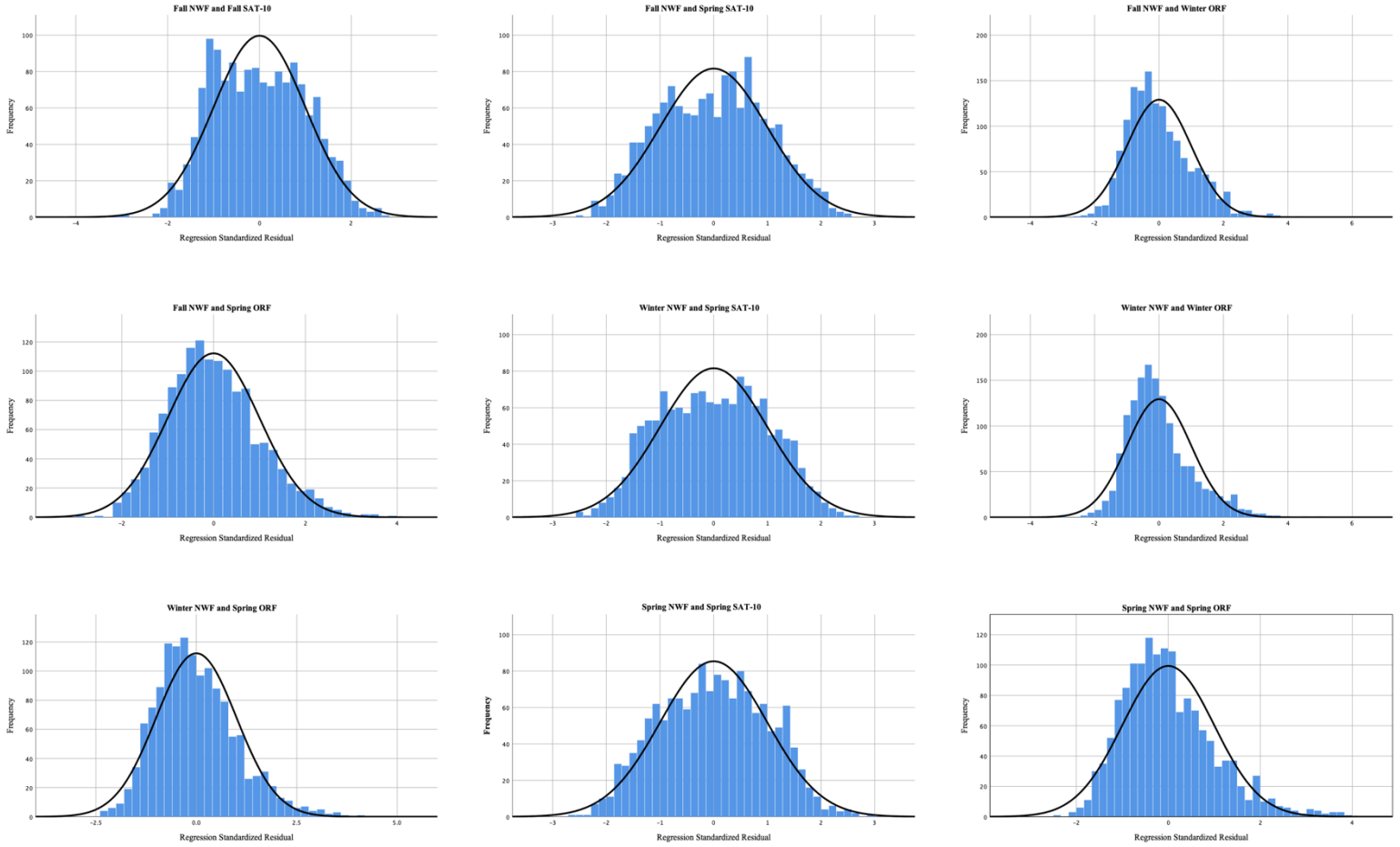
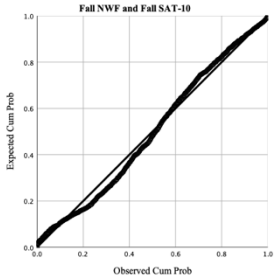
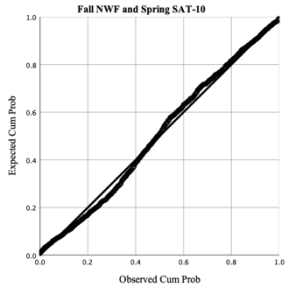
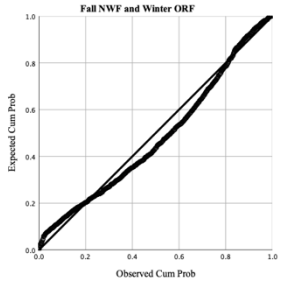


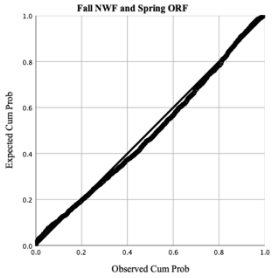
Figure 16

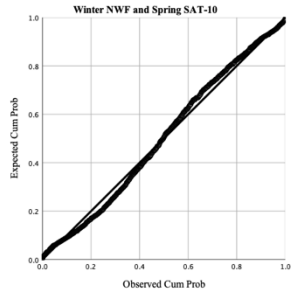
Normal P-P Plots of Outcomes Regressed on Screening Measures

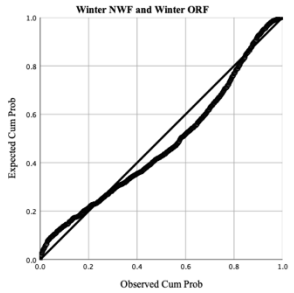


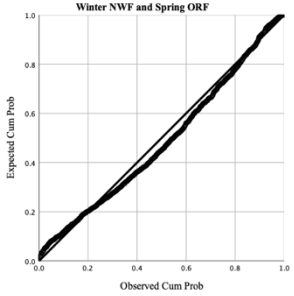


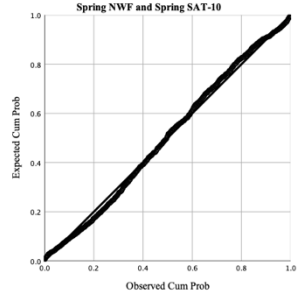


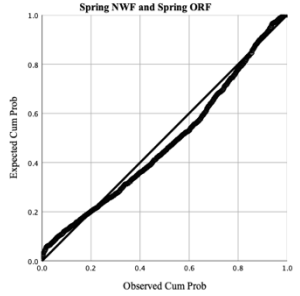












REFERENCES CITED

Baker, S. K., Fien, H., & Baker, D. L. (2010). Robust reading instruction in the early grades: Conceptual and practical issues in the integration and evaluation of tier 1 and tier 2 instructional supports. *Focus on Exceptional Children, 42*(9)*,* 1-20. https://doi.org/ 10.17161/foec.v42i9.6693

Bakker, A., Cai, J., English, L., Kaiser, G., Mesa, V., Dooren, W. V. (2019). Beyond small, medium, or large: Points of consideration when interpreting effect sizes. *Educational Studies in Mathematics, 102,* 1-8. https://doi.org/10.1007/s10649-019-09908-4

Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. M. (2015). *Evaluation of response to intervention practices for elementary school reading* (NCEE No. 2016-4000). Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. http://files.eric.ed.gov/fulltext/ED560820.pdf

Berkeley, S., Scanlon, D., Bailey, T. R., Sutton, J. C., & Sacco, D. M. (2020). A snapshot of RTI implementation a decade later: New picture, same story. *Journal of Learning Disabilities, 53*(5), 332-342. https://doi.org/10.1177/0022219420915867

Bossuyt, P. M., Reitsma, J. B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L., Lijmer, J.G., Moher, D., Rennie, D., de Vet, H.C.W., Kressel, H.Y., Rifai, N., Golub, R.M., Altman, D.G., Hooft, L., Korevaar, D.A., & Cohen, J.F. (2015). For the STARD group. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. https://doi.org/ 10.1136/bmj.h5527

Burke, M. D., & Hagan-Burke, S. (2007). Concurrent criterion-related validity of early literacy indicators for middle of first grade. *Assessment for Effective Intervention, 32*(2)*,* 66–77. https://doi.org/10.1177/15345084070320020401

Burns, M. K. (2012). Assessment research and school psychology: Introduction to the special series. *School Psychology Review, 41*(3)*,* 243-245. https://doi.org/10.1080/02796015.2012.12087505

Catts, H. W., Nielsen, D. C., Bridges, M. S., Liu, Y. S., & Bontempo, D. E. (2015). Early identification of reading disabilities within an RTI framework. *Journal of Learning Disabilities, 48*(3), 281-297. https://doi.org/10.1177/0022219413498115

Catts, H. W. & Petscher, Y. (2020). A cumulative risk and protection model of dyslexia. EdArXiv. https://doi.org/10.35542/osf.io/g57ph

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities, 42*(2), 163-176. https://doi.org/ 10.1177/0022219408326219

Choi, B. C. (1998). Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. *American Journal of Epidemiology, 148*(11), 1127-1132. https://doi.org/10.1093/oxfordjournals.aje.a009592

Clemens, N. H., Shapiro, E. S., & Thoemmes, F. (2011). Improving the efficacy of first grade reading screening: An investigation of word identification fluency with other early literacy indicators. *School Psychology Quarterly, 26*(2)*,* 231-244. https://doi.org/10.1037/a0025173

Cohen, J. (1988). Statistical power analysis for the behavioral Sciences (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. F., Korevaar, D. A., Altman, D. G., Brusn, D. E., Gatsonis, C. A., Hooft, L., … & Bossuyt, P. M. M. (2016). STARD 2015 guidelines for reporting diagnostic accuracy studies: Explanation and elaboration. *BJM Open, 6,* 1-17. https://doi.org/10.1136/bmjopen-2016-012799

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., Cho, E., & Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology, 102*(2), 327–340. https://doi.org/10.1037/a0018448

Compton, D. L., Gilbert, J. K., Jenkins, J. R., Fuchs, D., Fuchs, L. S., Cho, E., Barquero, L. A., & Bouton, B. (2012). Accelerating chronically unresponsive children to Tier 3 instruction: What level of data is necessary to ensure selection accuracy? *Journal of Learning Disabilities, 45*(3), 204-216. https://doi.org/10.1177/0022219412442151

Cummings, K. D., Dewey, E. N., Latimer, R. J., & Good, R. H., III. (2011). Pathways to word reading and decoding: The roles of automaticity and accuracy. *School Psychology Review, 40*(2)*,* 284–295. https://doi.org/ 10.1080/02796015.2011.12087718

Deeks, J. J. (2001). Systematic reviews of evaluations of diagnostic and screening tests. *BMJ, 323,* 157-162. https://doi.org/10.1136/bmj.323.7305.157

Deno, S. L. (1989). Curriculum-based measurement and special education services: A fundamental and direct relationship. In Shinn, M. R. (Ed.), Curriculum-based measurement: Assessing special children (pp. 1-17). The Guilford Press.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37(3),* 184-192. https://doi.org/10.1177/00224669030370030801

Espin, C., McMaster, K., Rose, S., & Wayman, M. (Eds.). (2012). *A measure of success: The influence of curriculum-based measurement on education*. University of Minnesota Press.

Eusebi, P. (2013). Diagnostic accuracy measures. *Cerebrovascular Diseases, 36*(4)*,* 267-272. https://doi.org/10.1159/000353863

Fien, H., Baker, S. K., Smolkowski, K., Mercier Smith, J. L., Kame'enui, E. J., & Beck, C. T. (2008). Using nonsense word fluency to predict reading proficiency in kindergarten through second grade for English learners and native English speakers. *School Psychology Review, 37*(3), 391-408. https://doi.org/ 10.1080/02796015.2008.12087885

Fien, H., Park, Y., Baker, S. K., Smith, J. L. M., Stoolmiller, M., & Kame'enui, E. J. (2010). An examination of the relation of nonsense word fluency initial status and gains to reading outcomes for beginning readers. *School Psychology Review, 39*(4)*,* 631–653. https://doi.org/ 10.1080/02796015.2010.12087747

Fien, H., Nelson, N. J., Smolkowski, K., Kosty, D. B., Pilger, M., Baker, S. K., Smith, J. L. M. (2020). A Conceptual replication study of the Enhanced Core Reading Instruction MTSS-reading model. *Exceptional Children,* Advance online publication. https://doi.org/10.1177/0014402920953763

Fien, H., Smith, J. L. M., Smolkowski, K., Baker, S. K., Nelson, N. J., Chaparro, E. (2015). An examination of the efficacy of a multitiered intervention on early reading outcomes for first grade students at risk for reading difficulties. *Journal of Learning Disabilities, 48*(6), 602-621. https://doi.org/10.1177/0022219414521664

Fleiss, J. L. (1981). *Statistical methods for rates and proportions (*2nd ed.). New York: Wiley.

Fuchs, D., Fuchs, L. S., & Compton, D. L. (2012). Smart RTI: A next generation approach to multilevel prevention. *Exceptional Children, 78*(3)*,* 263–279. https://doi.org/10.1177/001440291207800301

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children, 71*(1)*,* 7–21. https://doi.org/10.1177/001440290407100101

Fuchs, D., Mock, D., Morgan, P.L., & Young, C.L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research & Practice, 18*(3), 157-171. https://doi.org/10.1111/1540-5826.00072

Fuchs, L. S. & Vaughn, S. (2012). Responsiveness-to-intervention: A decade later. *Journal of Learning Disabilities, 45*(3), 195-203. https://doi.org/10.1177/0022219412442150

Gersten, R., Compton, D., Connor, C.M., Dimino, J., Santoro, L., Linan-Thompson, S., and Tilly, W.D. (2009). Assisting students struggling with reading: Response to Intervention and multi-tier intervention for reading in the primary grades. A practice guide. (NCEE 2009-4045). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/ncee/wwc/ publications/practiceguides/.

Gersten, R., Jayanthi, M, & Dimino, J. (2017). Too much, too soon? Unanswered questions from national response to intervention evaluation. *Exceptional Children, 83*(3), 244-254. https://doi.org/10.1177/0014402917692847

Gersten, R., Newman-Gonchar, R. A., Haymond, K. S., & Dimino, J. (2017). What is the evidence base to support reading interventions for improving student outcomes in grades 1–3? (REL 2017–271). Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast, U.S. Department of Education. http://ies.ed.gov/ncee/edlabs

Gilbert, J. K., Compton, D. L., Fuchs, D. & Fuchs, L. S. (2012). Early screening for risk of reading disabilities: Recommendations for four-step screening system. *Assessment of Effective Intervention, 38*(1), 6-14. https://doi.org/ 10.1177/1534508412451491

Goffreda, C. T., DiPerna, J. C., & Pedersen, J. A. (2009). Preventative screening for early readers: Predictive validity of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). *Psychology in the Schools, 46*(6)*,* 539–552. https://doi.org/10.1002/pits.20396

University of Oregon (2002). Dynamic Indicators of Basic Early Literacy Skills (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement. Available from http://dibels.uoregon.edu/

Greulich, L., Al Otaiba, S., Schatschneider, C., Wanzek, J., Ortiz, M., & Wagner, R. K. (2014). Understanding inadequate response to first-grade multi-tier intervention: Nomothetic and ideographic perspectives. *Learning Disability Quarterly, 37*(4), 204-217. https://doi.org/ 10.1177/0731948714526999

Grimes, D. A. & Schulz, K. F. (2005). Refining clinical diagnosis with likelihood ratios, *The Lancet, 9469*(365)*,* 1439-1514. https://doi.org/10.1016/S0140-6736(05)66422-7

Harn, B. A., Stoolmiller, M., & Chard, D. J. (2008). Measuring the dimensions of alphabetic principle on the reading development of first graders: The role of automaticity and unitization. *Journal of Learning Disabilities, 41*(2), 143–157. https://doi.org/10.1177/0022219407313585

Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Processing. *School Psychology Review, 32,* 541–556. https://doi.org/10.1080/02796015.2003.12086220

January, S. A., Ardoin, S. P., Christ, T. J., Eckert, T. L., & White, M. J. (2016). Evaluating the interpretations and use of curriculum-based measurement in reading and word lists for universal screening in first and second grade. *School Psychology Review, 45*(3)*,* 310-326. https://doi.org/10.17105/SPR45-3.310-326

January, S. A. & Klingbiel, D. A. (2020). Universal screening in grades K-2: A systematic review and meta-analysis of early reading curriculum-based measures. *Journal of School Psychology, 82,* 103-122. https://doi.org/10.1016/j.jsp.2020.08.007

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*(4), 582–600. https://doi.org/ 10.1080/02796015.2007.12087919

Jenkins, J. R., Schiller, E., Blackorby, J., Thayer, S. K., & Tilly, W. D. (2013). Responsiveness to intervention in reading: architecture and practices. *Learning Disability Quarterly, 36*(1), 36–46. https://doi.org/10.1177/0731948712464963

Johnson, E. S., Jenkins, J. R., & Petscher, Y. (2010). Improving the accuracy of a direct route screening process. *Assessment for Effective Intervention, 35*(3)*,* 131–142. https://doi.org/10.1177/1534508409348375

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). Can we improve the accuracy of screening instruments? *Learning Disabilities Research and Practice, 24*(4)*,* 174–185. https://doi.org/10.1111/j.1540-5826.2009.00291.x

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80*(4)*,* 437-447. https://doi.org/10.1037/0022-0663.80.4.437

Kame'enui, E. J. & Carnine, D. (1998). Effective teaching strategies that accommodate diverse learners. Upper Saddle River, NJ: Prentice Hall.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*(3)*,* 527-535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. (2006). Validation. In R. Brennan (Ed.), Educational Measurement (4th ed., pp. 17-64). Westport, CT: American Council on Education and Praeger.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review, 42*(4), 448-457. https://doi.org/10.1080/02796015.2013.12087465

Kent, P. & Hancock, M. J. (2016). Interpretation of dichotomous outcomes: Sensitivity, specificity, likelihood ratios, and pre-test and post-test probability. *Journal of Physiotherapy, 62*(4), 231-233. https://doi.org/ 10.1016/j.jphys.2016.08.008

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of School Psychology, 52*(4), 377-405. https://doi.org/10.1016/j.jsp.2014.06.002

Klingbeil, D. A., Van Norman, E. R., Nelson, P. M., & Birr, C. (2019). Interval likelihood ratios: Applications for gated screening in schools. *Journal of School Psychology, 76,* 107-123. https://doi.org/10.1016/j.jsp.2019.07.016

Leeflang, M. M. G., Bossuyt, P. M. M., & Irwig, L. (2009). Diagnostic test accuracy may vary with prevalence: Implications for evidence-based diagnosis. *Journal of Clinical Epidemiology, 62,* 5-12. https://doi.org/10.1016/j.jclinepi.2008.04.007

Malhotra, R., & Indrayan, A. A. (2010). A simple nomogram for sample size for estimating sensitivity and specificity of medical tests. *Indian Journal of Ophthalmology, 58,* 519–522. https://doi.org/10.4103/0301-4738.71699

McCardle, P., Scarborough, H. S., & Catts, H. W. (2001). Predicting, explaining, and preventing children's reading difficulties. *Learning Disabilities Research & Practice, 16*(4),230-239. https://doi.org/10.1111/0938-8982.00023

McGee, S. (2001). Evidence-based physical diagnosis. Philadelphia, PA: Elsevier.

Mellard, D. F., McKnight, M., & Woods, K. (2009). Response to intervention screening and progress monitoring practices in 41 local schools. *Learning Disabilities Research & Practice, 24*(4)*,* 186–195. https://doi.org/10.1111/j.1540-5826.2009.00292.x.

Messick, S. (1975). The standard problem: Meaning and values in measurement and education. *American Psychologist, 30*(10)*,* 955-966. https://doi.org/10.1037/0003-066X.30.10.955

Messick, S. (1989). Validity. In R. L. Linn (Ed.) Educational measurement (3rd ed. pp. 13-103.) New York, NY: American Council on Education and Macmillan.

National Center on Improving Literacy. (2020, 12). *State of Dyslexia.* https://improvingliteracy.org/state-of-dyslexia

National Center on Intensifying Intervention. (2021, 1). *Academic Screening Frequently Asked Questions (FAQ).* https://intensiveintervention.org/sites/default/files/NCII_AcademicScreening_FAQ_July2018.pdf

Nelson, J. M. (2008). Beyond correlational analysis of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS): A classification validity study. School Psychology Quarterly, 23, 542–552. https://doi.org/10.1037/a0013245

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine, 17*(8)*,* 857-872. https://doi.org/10.1002/(sici)1097-0258(19980430)17:8<857::aid-sim777>3.0.co;2-e

Pearson Education (2021, January 8). *SAT-10 Overview Flyer.* https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/sat10/sat10-overview-flyer.pdf

Pepe, M. S. (2003). The statistical evaluation of medical tests for classification and prediction. Oxford: New York.

Petscher, Y., Kim, Y., & Foorman, B. R. (2011). The importance of predictive power in early screening assessments: Implications for placement in the response to intervention framework. *Assessment for Effective Intervention, 36*(3), 158-166. https://doi.org/10.1177/1534508410396698

Reschly, A. L., Busch, T. W., Betts, J., Deno,S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*(6), 427-469. https://doi.org/10.1016/j.jsp.2009.07.001

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Not just speed reading: Accuracy of the DIBELS oral reading fluency measure of predicting high stakes third grade reading comprehension outcomes. *Journal of School Psychology, 46*(3), 343-366. https://doi.org/10.1016/j.jsp.2007.06.006

Rutjes, A. W.S., Reitsma, J. B., Nisio, M. D., Smidt, N., van Rijn, J. C., & Bossuyt, P. M. M. (2006). Evidence of bias and variation in diagnostic accuracy studies. *Canadian Medical Association Journal, 174*(4), 469-476. https://doi.org/10.1503/cmaj.050090

Samuels, C.A. (2011). An instructional approach expands its reach. *Education Week, 30*(22), 2-5. https://www.edweek.org/ew/articles/2011/03/02/22rti-overview.h30.html

Scarborough, H. S. (1998). Predicting the future achievement of second graders with reading disabilities: Contributions of phonemic awareness, verbal memory, rapid serial naming, and IQ. *Annals of Dyslexia, 48*, 115–36. https://doi.org/10.1007/s11881-998-0006-5

Schwartz, A. (2021, January 5). *Diagnostic accuracy calculator.* http://araw.mede.uic.edu/cgibin/testcalc.pl?DT=0&Dt=0&dT=0&dt=0&2x2=Compute

Shapiro, E., Solari, E., & Petscher, Y. (2008). Use of an assessment of reading comprehension in addition to oral reading fluency on the state high stakes assessment for students in Grades 3 through 5. *Journal of Learning and Individual Differences, 18*(3)*,* 316-328. https://doi.org/10.1016/j.lindif.2008.03.002

Shinn, M. R. & Bamonto, S. (1998). Advanced applications of curriculum-based measurement: "Big Ideas" and avoiding confusion. In M. R. Shinn (Ed), *Advanced applications of curriculum-based measurement* (pp. 1-31). The Guildford Press.

Silberglitt, B. & Hintze, J. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment, 23(4),* 304-325. https://doi.org/10.1177/073428290502300402

Silva, M. R., Collier-Meek, M. A., Codding, R. S., Kleinert, W. L., & Feinberg, A. (2020). Data collection and analysis in response-to-intervention: A survey of school psychologists. *Contemporary School Psychology, 24*(1), https://doi.org/10.1007/s40688-020-00280-2

Smith, J. L. M., Nelson, N. J., Smolkowski, K., Baker, S. K., Fien, H. & Kosty, D. (2016). Examining the efficacy of a multitiered intervention for at-risk readers in grade 1. Elementary School Journal, 116(4), 549-573. https://doi.org/10.1086/686249

Smolkowski, K. & Cummings, K. D. (2015). Evaluation of diagnostic systems: The selection of students at risk of academic difficulties. *Assessment for Effective Intervention, 41*(1), 41-54. https://doi.org/10.1177/1534508415590386.

Smolkowski, K. & Cummings, K. D. (2016). Evaluation of the DIBELS (Sixth Edition) diagnostic system for the selection of native and proficient English speakers at risk of reading difficulties. *Journal of Psychoeducational Assessment, 34*(2), 103-118. https://doi.org/10.1177/0734282915589017

Smolkowski K., Cummings K., Strycker L. (2016). An introduction to the statistical evaluation of fluency measures with signal detection theory. In K. Cummings & Y. Petscher (Eds.), *The Fluency Construct* (pp. 187-221). Springer. https://doi.org/10.1007/978-1-4939-2803-3_8

Speece, D. L. (2005). Hitting the moving target known as reading development: Some thoughts on screening children for secondary interventions. *Journal of Learning Disabilities, 38*(6), 487-493. https://doi.org/10.1177/00222194050380060301

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285–1293. https://doi.org/10.1126/science.3287615

Swets, J. A. (1996). Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers. Hillsdale: Lawrence Erlbaum Associates.

Tindal, G. (1989). Evaluating the effectiveness of educational programs at the systems level using curriculum-based measurement. In M. Shinn (Ed.), Curriculum-based assessment: Assessing special children (pp. 202–238). New York, NY: Guilford Press.

Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. ISRN Education (International Scholarly Research Network), 2013, 1– 29. https://doi.org/10.1155/2013/958530

Toste, J. R., Compton, D. L., Fuchs, D., Fuchs, L. S., Gilbert, J. K., Cho, E., Barquero, L. A., & Bouton, B. D. (2014). Understanding unresponsiveness to Tier 2 reading intervention: Exploring the classification and profiles of adequate and inadequate responders in first grade. *Learning Disability Quarterly, 37*(4), 192-203. https://doi.org/10.1177/0731948713518336

Van Norman, E. R., Klingbeil, D. A., & Nelson, P. M. (2017). Posttest probabilities: An empirical demonstration of their use in evaluating the performance of universal screening measures across settings. *School Psychology Review, 46*(4), 349–362. https://doi.org/10.17105/SPR-2017-0046.V46-4

VanDerHeyden, A. M. (2011). Technical adequacy of response to intervention decisions. *Exceptional Children, 77*(3), 335–350. https://doi.org/10.1177/001440291107700305

VanDerHeyden, A. M. (2013). Universal screening may not be for everyone: Using a threshold model as a smarter way to determine risk. *School Psychology Review, 42*(4), 402–414. https://doi.org/10.1080/02796015.2013.12087462

VanDerHeyden, A. M., & Burns, M. K. (2018). Improving decision making in school psychology: Making a difference in the lives of students, not just a prediction about their lives. *School Psychology Review, 47*(4), 385-395. https://doi.org/10.17105/SPR-2018-0042.V47-4

VanDerHeyden, A. M., Burns, M. K., & Bonifay, W. (2018). Is more screening better? The relationship between frequent screening, accurate decisions, and reading proficiency. *School Psychology Review, 47*(1), 62-82. https://doi.org/10.17105/SPR-2017-0017.V47-1

Wanzek, J., Vaughn, S., Scammacca, N., Gatlin, B., Walker, M. A., & Capin, P. (2016). Meta-analyses of the effects of Tier 2 type reading interventions in grades K-3. *Educational Psychology Review, 28,* 551-576. https://doi.org/10.1007/s10648-015-9321-7

Whiting, P., Rutjes, A. W. S., Reitsma, J., Glas, A. S., Bossuyt, P. M. M., & Kleijnen, J. (2004). Sources of variation and bias in studies of diagnostic accuracy: A systematic review. *Annals of Internal Medicine,140(3),* 189-203. https://doi.org/10.7326/0003-4819-140-3-200402030-00010

Whiting, P.F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J.B., Leeflang, M. M., Sterne, J. A. C., & Bossuyt, P. M. M. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine, 155*(8), 529-536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

Whiting, P.F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., & the QUADRAS-2 Steering Group (2013). A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *Journal of Clinical Epidemiology, 66*(10)*,* 1093-1104. https://doi.org/10.1016/j.jclinepi.2013.05.014

Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education, 31*(6), 412-422. https://doi.org/10.1177/0741932508327463