

Examining the Predictive Validity of Universal Screening Measures of Reading in Grades 4-5

Patrick Gipson

A dissertation

submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Kristen N. Missall, Chair

Carol A. Davis

Elizabeth A. Sanders

Erin M. Ingvalson

Program Authorized to Offer Degree:

College of Education

©Copyright 2026

Patrick Gipson

University of Washington

Abstract

Examining the Predictive Validity of Universal Screening Measures of Reading in Grades 4-5

Patrick Gipson

Chair of the Supervisory Committee:

Kristen N. Missall, Ph.D.

College of Education

Universal screening of reading is an increasingly common practice within schools (Silverman et al., 2021), and it's often considered a core component of a multi-tiered system of supports (MTSS) framework (Compton et al., 2006; Jenkins et al., 2007). Universal screening of reading can be used to identify students that may benefit from additional supports, assess key academic skills, allocate resources effectively, and predict performance on state assessments (Ardoin & Christ, 2008; January & Klingbeil, 2020; Jenkins et al., 2007; Missall et al., 2021). Universal screening of reading measures vary in terms of length, content, and psychometric quality. The current study compared the predictive validity of three universal screening measures of reading: prior year Smarter Balanced Assessment – English Language Arts (SBA-ELA), iReady Reading Diagnostic, and Dynamic Indicators of Basic Early Literacy – Oral Reading Fluency (DIBELS-ORF). Additionally, the moderating effects of grade level and learning

assistance on predictive utility were tested. The results indicated that all three measures significantly and substantially predicted end-of-year SBA-ELA scores, but that iReady Reading was the strongest predictor – albeit only slightly better than previous year SBA-ELA scores. In contrast, DIBELS-ORF did not contribute meaningfully to explaining variation in end-of-year literacy scores beyond the other two measures. Limitations and implications for data-based decision making and universal screening for reading are discussed.

Keywords: universal screening of reading, data-based decision making, iReady Reading, DIBELS-ORF, SBA-ELA, multi-tiered system of supports

Acknowledgements

The work represented in this dissertation is a testament to not only my efforts, but to the love, dedication, and sacrifice of the community that I have around me. First, I'd like to give honor to God for His unending love, unfailing grace, and unwavering protection through the peaks and valleys of life. Second, I'd like to thank my family for being a constant source of love, encouragement, and laughter. I'd like to give a special thank you to my mother, Pearl, who is the personification of resilience, love, kindness, and faith. Your words of wisdom and encouragement have made this journey possible. I would also like to thank my aunt, Dr. Linda Jones, who passed shortly before the completion of this project. Her fierce devotion to family, faith, and service to others has been a constant inspiration.

In addition to the support of my family, I've been fortunate to have a thoughtful, knowledgeable, kind, and supportive committee. I'd like to thank Dr. Carol Davis for her mentorship, generosity, and her passion for building effective systems. I'd like to thank Dr. Erin Ingvalson for her thoughtful feedback, for sharing her expertise on early language screening, and for her dedication to supporting the literacy and language development of all children. I'd like to extend a special thank you to Dr. Elizabeth Sanders, whose immense knowledge is matched only by her kindness, generosity, and passion for helping her students succeed. I'm constantly inspired by her passion for guiding and mentoring future generations of scholars. And it is with deep gratitude that I acknowledge my mentor, teacher, and committee chair, Dr. Kristen Missall. I could not have asked for a kinder, more dedicated, more giving human being to guide me on this journey. As I reflect on the countless hours you have spent meeting with me, reading my work, providing feedback, advocating on my behalf, helping me develop my professional identity, and encouraging me, I feel both grateful and a sense of responsibility. I feel a

responsibility to show up for my future trainees and advisees the way you've shown up for me. I will be forever thankful for your support.

Additionally, I'd like to thank my wonderful friends and cohort mates: Jason, Josh, Emalise, Katrina, Lauren, and Bella. It has been an honor to go on this journey with you all, and I'm so excited to see the amazing work that you will do in the future. I'd also like to thank my school district partners, without whom this project would not have been possible. Last, I'd like to thank the woman of my dreams and my answered prayers, my wife Kayla.

Kayla, I do not know what I did to deserve such a loving, caring, intelligent, strong, and devoted partner. Thank you for bringing laughter when life felt somber; thank you for bringing peace when life felt chaotic; and thank you for being a constant when life felt uncertain. It has been such a blessing to build a life with you, Sunita, and Kong. I love you beyond forever.

Table of Contents

List of Figures	iv
List of Tables	v
Chapter 1: Introduction	1
The National Assessment of Educational Progress	1
How does NAEP Measure Reading?	2
Overall Reading Trends of NAEP Across Decades	3
Subgroup Analysis and Trends	4
<i>Race</i>	4
<i>Language Proficiency</i>	5
<i>Ability</i>	6
<i>Intersection of Ability and Language Proficiency</i>	7
<i>Family Income</i>	8
<i>Geography</i>	8
<i>Gender</i>	10
Equity Implications of National Reading Trends	10
Limitations of NAEP Reading	11
Using Universal Screening of Reading to Identify and Address Reading Difficulty	14
<i>Efforts to Prevent and Remediate Reading Difficulties in Washington State</i>	15
<i>What is Universal Screening for Reading?</i>	15
<i>How does Universal Screening Address Limitations of NAEP?</i>	15
Chapter 2: Literature Review	17
Determining the Psychometric Quality of Universal Screeners of Reading	17
Data-Based Decision Making Using Universal Screeners of Reading	19
Learning Assistance Program (LAP) Services	20
Universal Screening of Reading in Grades 4 and 5	20
Reading Skills Assessed during Universal Screening	21
<i>Word Recognition</i>	22
<i>Oral Reading Fluency</i>	24
<i>Vocabulary</i>	26

<i>Comprehension</i>	28
Types of Universal Screening Measures of Reading	29
<i>Curriculum-Based Measures</i>	29
<i>CBM-R and State Assessments</i>	31
<i>Computer Adaptive Testing</i>	32
<i>State Assessments</i>	33
Commonly Used Measures for Universal Screening of Reading	35
<i>iReady Reading Diagnostic</i>	35
<i>Dynamic Indicators of Basic Early Literacy Skills, 8th Edition Oral Reading Fluency (DIBELS-ORF)</i>	36
Present Study	37
<i>Research Questions</i>	39
Chapter 3: Method	40
Sample	40
Measures	40
<i>Smarter Balanced Assessment – English Language Arts (SBA-ELA)</i>	40
<i>iReady Reading Diagnostic Assessment</i>	42
<i>Dynamic Indicators of Basic Early Literacy – Oral Reading Fluency (DIBELS-ORF)</i>	45
Data Analysis Plan	46
<i>General Modeling Approach</i>	46
<i>Research Questions 1 and 2: Predictive Utility of Universal Screeners and Moderators</i>	47
<i>Research Question 3: Unique Predictive Utility of Universal Screeners</i>	48
Chapter 4: Results	49
Descriptive Statistics	49
Intraclass Correlations	50
Research Question 1: Direct Predictive Utility of Reading Screeners	50
Research Question 2: Grade and LAP Effects on Screener Relations	52
Research Question 3: Unique Predictive Utility of Reading Screeners	52
Chapter 5: Discussion	54
Research Question 1: Direct Predictive Utility of Reading Screeners	55
Research Question 3: Unique Predictive Utility of Reading Screeners	56

Research Question 2: Grade and LAP Effects on Screener Relations	58
Limitations and Future Directions	60
<i>Threats to Internal Validity</i>	60
<i>Threats to External Validity</i>	60
Conclusion	61
References	63

List of Figures

Figure 1: Model-Predicted Mean Relations between Reading Screeners and Spring 2024 SBA- ELA (Model 1).....	104
Figure 2: Model-Predicted Relations between Reading Screeners and Spring 2024 SBA-ELA by Grade and LAP Status (Model 1).....	105

List of Tables

Table 1: Sample Demographic Characteristics	99
Table 2: SBA-ELA Reading Levels and Skill Expectations for Students in Grades 4-5	100
Table 3: Sample Descriptive Statistics and Zero-Order Correlations	101
Table 4: Predicting Spring 2024 SBA-ELA with Each Reading Screener in Isolation (Model 1)	102
Table 5: Predicting Spring 2024 SBA-ELA with Combinations of Reading Screeners (Model 2)	103

Chapter 1: Introduction

Reading achievement in the United States (U.S.) has been a widely publicized and debated topic for decades. The popularity of this topic is linked to the academic, social, and vocational outcomes associated with reading. Researchers have found that reading skills predict performance in other content areas such as math (Björn et al., 2016; Purpura et al., 2011; Vilenius-Tuohimaa et al., 2008), science (Reed et al., 2017; Zhu, 2022), writing (Cabell et al., 2022), and social studies (Ozensoy, 2021; Wanzek et al., 2018). Additionally, reading skills predict graduation rates and quality of later job opportunities (Smart et al., 2017).

A review of state and national reading outcomes shows that many students in the U.S. are struggling to read. Results from the 2025 administration of the Smarter Balanced Assessment English Language Arts (SBA-ELA), a state assessment administered to Washington students in grades 3 through 8 and grade 10, show that nearly 50% of students received scores below grade level expectations (Office of the Superintendent of Public Instruction [OSPI], 2025). This percentage is even higher when examined by subgroups, with students of color, multilingual learners (MLL), students with identified disabilities, and students with limited economic resources scoring lower on the SBA-ELA than their peers. These state-level reading results mirror national reading results as described by the National Assessment of Educational Progress (NAEP).

The National Assessment of Educational Progress

At its inception, NAEP was created to assess and monitor the academic knowledge of students in the U.S. across time and across geographic regions (Buckendahl et al., 2009a; Buckendahl et al., 2009b; Jones, 1996; Lane et al., 2009). Use of NAEP data has expanded to analyzing and monitoring the performance of subgroups of students over time (Johnson, 1992;

Wells et al., 2009). There are two forms of the NAEP test: the main NAEP and the long-term trend NAEP (David & Marchant, 2015). The item content of the main NAEP is updated frequently to align with curricula and educational standards followed in the U.S. The item content of the long-term trend NAEP remains relatively consistent to facilitate comparisons of NAEP results over time (Fahle et al., 2019; Johnson, 1992; Rampey et al., 2009; Reilly et al., 2019). Although, updates to administration procedures, adjustments to outdated questions, and the inclusion of accommodations were made to the long-term trend NAEP in 2004 (Rampey et al., 2009).

How does NAEP Measure Reading?

NAEP was first administered in 1969 in response to concerns about the academic achievement of U.S. students (Fahle et al., 2019; National Academies of Sciences, Engineering, and Medicine, 2017). The first testing year attempted to measure knowledge in the areas of citizenship, writing, and science. The reading NAEP was first administered in the 1970s (Fahle et al., 2019), measuring academic skills through multiple choice, short response, and extended response formats (David & Marchant, 2015; Jones, 1996; Rampey et al., 2009). The main goal of the NAEP reading test is to measure how well a student uses background and vocabulary knowledge to understand what they read (National Assessment Governing Board [NAGB], 2022). Specifically, the reading portions of NAEP measure the ability to locate information in text, make inferences, determine the main idea, and use evidence to support conclusions (Rampey et al., 2009; Reardon et al., 2012).

Items on the reading portion of the NAEP have been administered digitally since 2017 (NAGB, 2022). Additionally, the reading NAEP tasks involve comparing, evaluating, and explaining the relationships between texts (Polat et al., 2016; Rindermann & Thompson, 2013).

The NAEP reading test includes three subscales: Reading for Literacy Experience, Reading for Information, and Reading to Perform a Task (Wells et al., 2009).

The NAEP's reading framework was updated in 2009 (National Academies of Sciences, Engineering, and Medicine, 2017), including more items that measure vocabulary knowledge (NAGB, 2022). Additional modifications to the grade 12 version of the reading NAEP have been made to assess readiness for post-secondary education (NAGB, 2022). The updated framework also includes the addition of poetry for grade 4, a more systematic assessment of vocabulary, and a greater variety of texts (NAGB, 2022).

While participating in the reading NAEP, students are required to read and answer questions about a variety of texts, including poems, reports, stories, and advertisements (NAGB, 2022; Rampey et al., 2009; Reilly et al., 2019). The NAEP categorizes texts as either literary or informational (NAGB, 2022). The overall assessment contains 36 to 40 passages, with each individual student reading between 8 and 15 passages (Rampey et al., 2009). Johnson (1992) estimated that fourth-grade students spend about 60% of their testing time answering open-ended questions, while students in grades 8 and 12 dedicate about 75% of their time completing these tasks. More recently, the time is estimated to be slightly lower, with longer responses estimated to take up 50% of the assessment time for grade 4 students and 60% of the assessment time for students in grades 8 and 12 (NAGB, 2022).

Overall Reading Trends of NAEP Across Decades

According to the 2024 NAEP results, 37% of fourth-grade students and 30% of eighth-grade students received reading scores below the Basic proficiency level (National Center for Educational Statistics [NCES], n.d.). Approximately 37% of grade 12 students scored at or above the “proficient” level on the 2009 NAEP, which was the last year that scores were available for

this grade level (Reardon et al., 2012). Additionally, racial, ethnic, and gender-based score differences on NAEP can be seen across age groups (Reardon et al., 2012).

Subgroup Analysis and Trends

NAEP uses demographic questionnaires to obtain information on a students' ethnicity, family income, resources, and attitudes about school (Fahle et al., 2019; Johnson, 1992; Spees et al., 2016). This demographic information has been used to examine performance within and across subgroups of students (Fahle et al., 2019). David and Marchant (2015) found that among grade 4 students, the largest score gaps were seen for Black students, Latine students, and students with limited socioeconomic resources. Additionally, score differences between students with and without disabilities have been observed over time (NCES, n.d.).

Race

Since its inception, White students have consistently scored higher than their Black and Latine peers on the NAEP (David & Marchant, 2015; NCES, n.d.). Over the course of the COVID-19 pandemic, significant score decreases for Black, Latine, and White 9-year-olds students were seen (NCES, n.d.). While score decreases within racial groups have occurred, score differences between these groups have persisted (NCES, n.d.).

In 1971, the gap between Black and White students in reading was 44 points (David & Marchant, 2015). This gap reduced substantially in the 1970s and 1980s (Reardon et al., 2012). Between 1992 and 2007, the reading scores for both Black and White students increased; however, score differences between these groups continued, with Black students receiving scores 26 and 27 points lower than White students in 2004 and 2007, respectively (Vanneman et al., 2009). Researchers have also examined these score differences at the state level. For example,

Vanneman and colleagues (2009) identified score differences between Black and White students in 44 states for 9-year-olds and 41 states for 13-year-olds.

Analysis of the 2008 NAEP reading results indicated that Black students demonstrated faster growth than White students when compared to the 1971 reading results (Rampey et al., 2009). Latine students' reading scores followed a similar pattern, except for scores for students at age 13, which showed no significant difference between 1975 and 2008 (Rampey et al., 2009). Score differences between White students and Black and Latine students in 2008 narrowed since 1971 and 1975, respectively (Rampey et al., 2009). However, between 2011 and 2013 score differences between White and Black students increased (David & Marchant, 2015). Score differences between Latine and White students have followed a similar pattern.

The score difference between Latine and White students was widest in 1975, decreased by the 1980s, increased slightly in the 1990s, and decreased again in the 2000s (Readon et al., 2012). Hemphill and Vanneman (2011) examined NAEP reading data from 1992 to 2009, finding that the score differences for these two groups did not change significantly during this time at the national level. However, significant changes in scores were seen at the state level. It should be noted, however, that data from some states were not available, as some states did not have a high enough percentage of Latine students to analyze (Hemphill & Vanneman, 2011).

Language Proficiency

After examining NAEP reading data between 2003 and 2011, Polat and colleagues (2016) found that multilingual (ML) students in grades 4 and 8 received lower reading scores than their monolingual peers in every year examined. While ML students showed gains in between 2003 and 2005, these score increases appear to have diminished by 2011 (Polat et al., 2016). Monolingual learners' performance was relatively stable, although positive gains were

demonstrated during some years (Polat et al., 2016). In contrast, the performance of ML students showed more variability, with both positive and negative changes seen in the year range examined (Polat et al., 2016).

Between 2003 and 2015, the score differences between monolingual and ML students decreased but remained significant (Goodrich et al., 2021). Larger reading score differences between monolingual and ML fourth graders were seen in states with a higher number of Spanish speaking students than those with fewer Spanish speaking students; however, the states with a higher number of Spanish speaking students saw larger reductions of these score differences over time (Goodrich et al., 2021). Significant improvements in reading score performance have been seen among ML and monolingual African American students (Polat et al., 2016).

Ability

Between 1998 and 2022, students with disabilities received lower NAEP reading scores than their typically developing peers (NCES, n.d.). For grade 4 students, there was a 40-point score difference between these groups in both 1998 and 2022. The size of the score difference has varied in the intervening years; however, the difference was largest in 2000, which saw a 48-point difference between these two groups. Between 1998 and 2022, the second largest score difference across these groups was seen in 2019. This score difference was 42 points (NCES, n.d.). As of 2019, 70% of fourth-grade students with disabilities received scores in the Below Basic range, 18% received scores at Basic, 9% at Proficient, and 2% at Advanced (NCES, n.d.).

For eighth-grade students, score differences have been documented during every testing year since 1998 (NCES, n.d.). The score gap decreased slightly from 40 to 36 points between 1998 and 2022. According to the most recent NAEP reading data, 64% of grade 8 students

scored in the Below Basic range, 27% within the Basic range, 8% within the Proficient range, and 1% within the Advanced range (NCES, n.d.). Data from grade 12 students show a similar pattern, with score gaps shrinking from 47 points to 39 points between 1998 and 2019, the last year that data were available for this group (NCES, n.d.). For grade 12 students with disabilities, 64% received scores within the Below Basic range, 25% within the Basic range, 10% within the Proficient range, and 1% within the Advanced range (NCES, n.d.).

Several scholars have argued that the inclusion and exclusion criteria for students with disabilities varies across states, making it difficult to identify trends for this population (Braun et al., 2010; Lane et al., 2009). Students with disabilities can be excluded from NAEP if their cognitive abilities limit meaningful participation in the assessment, their individualized education program (IEP) team has determined participation would be inappropriate, or the accommodations listed in the student's IEP are not permitted by the NAEP (Vanneman et al., 2009). It should be noted that accommodations for students with disabilities and ML students were not permitted on the NAEP until 1996 and 2004, respectively (National Academies of Sciences, Engineering, and Medicine, 2017; Vanneman et al., 2009).

Intersection of Ability and Language Proficiency

Starting in 2015, NAEP began reporting the performance of ML students separately (Kangas et al., 2024). While investigating NAEP data from 2009 to 2019, Kangas and colleagues (2024) found that the percentage of students scoring in the “Below Basic” range on NAEP was 88% to 91% for ML students with disabilities, 63% to 72% for ML students without disabilities, 57% to 76% for students who formerly qualified for ML supports and have disabilities, and 21% to 33% of students who formerly met ML status and do not have disabilities. ML students with

disabilities tend to receive lower NAEP scores than those without disabilities, and these score differences have grown over time (Cooc, 2023).

Family Income

NAEP uses eligibility for free and reduced lunch as a proxy for family income (Vanneman et al., 2009). Achievement gaps for students of limited socioeconomic resources have increased over time and across racial groups (David & Marchant, 2015). As of 2022, students who qualified for free and reduced lunch tended to score lower than their peers who did not qualify for this program. Specifically, score differences in fourth and eighth grade were 28 and 23 points, respectively (NCES, n.d.). Socioeconomic status (SES) had more of an impact on the reading scores of White students than for Black and Latine students (David & Marchant, 2015). SES showed less of an impact on scores for White female students than for their male peers, while Black and Latine male students showed less SES related score differences in reading than Black and Latine female students (David & Marchant, 2015).

Research indicates that students who formerly received ML supports, qualified for free and reduced lunch, and had identified disabilities outperformed peers from similar socioeconomic and ability backgrounds who currently receive ML supports (Kangas et al., 2024). Additionally, ML students with identified disabilities who did not qualify for free and reduced lunch received higher scores than their peers from lower SES contexts (Kangas et al., 2024).

Geography

Regional score reporting categories are aligned with regions reported in the census, using the labels of Northeast, South, Midwest, and West (Polat et al., 2016). State-level reporting of NAEP data was first authorized in 1990 (Johnson, 1992), with state-level reading results first

reported in 1992 (Fahle et al., 2019). Between 2020 and 2022, significant decreases in fourth grade reading scores occurred across geographic locations, except for Western states (NCES, n.d.). Grade 4 students with disabilities in the Northeast received scores higher than the national average for this group in 2019, while students with disabilities in the West received lower scores than the national average (NCES, n.d.). A different pattern was observed for students with disabilities in eighth grade. In 2019, students with disabilities in the Northeast, South, and the West received significantly higher scores when compared to the 2003 testing year. Further, students from this group in both Northeastern and Midwestern states scored higher than the national average score for students with disabilities (NCES, n.d.).

For grade 4 students, 11 states reported increasing scores for both White and Latine students, with New York and New Jersey showing decreasing score differences between these groups (Hemphill & Vanneman, 2011). However, the gap widened in Colorado, between 1992 and 2009, and in Indiana, between 2002 and 2009 (Hemphill & Vanneman, 2011). The score difference for Latine and White eighth-grade students was narrower in seven states over the same period (Hemphill & Vanneman, 2011). None of the states included in the study demonstrated increasing score differences for eighth-grade students.

Grade 4 ML students with disabilities in both urban and suburban settings received higher NAEP reading scores than their peers in rural areas, while higher scores were seen in suburban areas compared to urban areas for eighth-grade students (Kangas et al., 2024). It is difficult, however, to make within-state comparisons as state-level data are the lowest, large-scale geographical reporting category (Spees et al., 2016), and sampling procedures for NAEP have led to underrepresentation of culturally and linguistically diverse groups in certain states (Buckendahl et al., 2009a).

Gender

After examining NAEP data from 1988 to 2015, Reilly and colleagues (2019) found that girls tended to outperform boys on the reading NAEP. This pattern persisted across grades, years (Reilly et al., 2019), and language (Polat et al., 2016). Additionally, boys were overrepresented in the score categories below proficiency, while more girls scored at or above proficiency (Reilly et al., 2019). In 2022, girls tended to outperform boys by 6, 8, and 12 points in grades 4, 8, and 12, respectively (NCES, n.d.).

Equity Implications of National Reading Trends

NAEP provides metrics for examining score equity, or the extent to which test scores can be adequately interpreted across sub-groups (Wells et al., 2009). Specifically, Wells and colleagues (2009) found that NAEP's classification categories remained relatively consistent across several states regardless of whether the scores were examined separately by state or by a national sample, and they may be used to compare performance between subgroups (Wells et al. 2009). However, while NAEP reporting procedures can help to identify the presence of score differences between groups, there is no information about the causes of such differences, which could include demographic changes in the population; instructional factors; socioeconomic factors; school conditions; or the influence of national, state, or school policies (David & Marchant, 2015; Lee, 2002; Vanneman et al., 2009). When score differences are examined without attending to these nuances, it can be tempting to make causal claims that are not supported by the data reported, which in turn can have ramifications for educators, families, policy makers, and students.

Classroom level factors, such as student-teacher ratios and access to qualified teachers may contribute to group differences observed on NAEP. Lee and Reeves (2012) examined

NAEP data from 1990 and 2009, finding that higher teacher and student ratios were associated with widening score gaps between White and minoritized students. Additionally, studies suggest that minoritized students are more likely to be taught by novice teachers (Clotfelter et al., 2005; Hanselman, 2019), although Hanselman (2019) found that differential access to quality teachers was not a major contributor to score differences.

Systemic barriers also impact access to effective practices for minoritized populations (Kangas et al., 2024), as well as participation in NAEP. Moreover, the inappropriate exclusion of both ML learners and students with disabilities has likely introduced bias to NAEP assessment results (Braun et al., 2010). Additionally, David and Marchant (2015) reported that SES-related score differences increased across gender and racial groups from 2000 to 2013, highlighting that culturally and linguistically diverse students may be more likely to encounter socioeconomic barriers (Aud et al., 2010), which could negatively impact academic achievement (David & Marchant, 2015).

Analyses of NAEP data have suggested that educators may need more training in supporting the literacy of ML learners with disabilities (Kangas et al., 2024), although these data do not provide information for designing effective instruction. Researchers have also critiqued how NAEP data show score differences between groups. While NAEP data provide a useful metric for examining reading performance within and between groups, the utility of this assessment has limits.

Limitations of NAEP Reading

Researchers have argued that NAEP data allow for state-level comparisons (Spees et al., 2016; Wells et al., 2009). However, these types of comparisons are difficult to make for certain subgroups of students. For example, it is difficult to compare performance of students with

disabilities across states because inclusion criteria and use of accommodations have differed by state historically (Buckendahl et al., 2009a). It is also difficult to make within state comparisons due to how geographic data are reported (Spees et al., 2016).

Further, NAEP may underestimate the academic skill of twelfth graders (Jones, 1996; Lane et al., 2009), as many of these students may be more focused on graduating than on the assessment (Jones, 1996). Additionally, the small number of difficult items may limit NAEP's ability to differentiate students scoring in the advanced range from those scoring within the proficient range (Jones, 1996).

Some scholars have argued that the inadequate representation of certain subgroups undermines the validity of the interpretations that can be drawn from NAEP data (Buckendahl et al., 2009a; Lane et al., 2009). Additionally, different cohorts of students are assessed during each cycle of the NAEP, making it difficult to interpret changes in scores across time (Reardon et al., 2012). Further, NAEP's description of appropriate use makes it challenging for the test developers to predict how its results may be misinterpreted and misused (Buckendahl et al., 2009a; Buckendahl et al., 2009b).

Buckendahl and colleagues (2009b) conducted a focus group of state NAEP coordinators, finding that many coordinators noted that stakeholders often compared scores across states without understanding the limitations of such comparisons, tended to interpret small changes in NAEP scores as significant, and often misinterpreted NAEP results. This is consistent with arguments that NAEP has been interpreted in ways that are inconsistent with the available validity evidence for this assessment (Buckendahl et al., 2009a). For example, although NAEP is often used to monitor achievement of culturally and linguistically diverse student groups, its

sampling procedures have led scholars to question the precision of NAEP for this purpose (Barron & Koretz, 1995; Buckendahl et al., 2009a).

Barron and Koretz (1995) argued that the sample size of minoritized students participating in NAEP tends to be quite small, increasing the amount of error associated with the assessment. Interpretation of results across groups is also complicated by the multi-staged nature of NAEP's sampling procedures, in which no student receives every item on the assessment (Buckendahl et al., 2009a). Barron and Koretz (1995) also identified psychometric issues with the use of age-based samples for the long-term trend NAEP and grade-based samples for the main NAEP. For example, the researchers found that when long-term trend scores of Black and Latine students were examined using grade instead of age, the performance of fourth-grade students from these groups appeared lower, while eighth-grade performance appeared higher (Barron & Koretz, 1995). Others have argued that the criteria for meeting the NAEP proficiency levels are unrealistically high (Buckendahl et al., 2009a; Fahle et al., 2019), encouraging misleading beliefs about overall student achievement in the U.S. The variability of exclusion rates for both ML learners and students with disabilities also makes it difficult to compare results across states (Braun et al., 2010), which is often how NAEP results are interpreted (Buckendahl et al., 2009a).

Several factors limit the NAEP's utility for informing instruction. First, while the NAEP reading test results provide a picture of national achievement trends, they do not provide information about how best to support the reading development of students in the U.S. (NAGB, 2022). Second, the NAEP does not assess some important aspects foundational to conventional reading, such as phonemic awareness, phonics, fluency, and language comprehension (NAGB,

2022). Third, NAEP scores are not reported at either the student or school level, placing limitations on how well these scores can be used to inform instruction.

Using Universal Screening of Reading to Identify and Address Reading Difficulty

Decades of NAEP reading results highlight that many U.S. students experience reading difficulties, creating a social, political, and educational imperative to address these challenges. Further, state-level reading scores mirror that of national trends. According to the Minnesota Department of Education, nearly 51% of students received below grade level scores on the Minnesota Comprehensive Assessment (MCA) in 2024 (Minnesota Department of Education, n.d.). Additionally, four years' worth of assessment results reveal score differences based on race, language, ability, and socio-economic status.

In Florida, which uses the Florida Assessment of Student Thinking (FAST), approximately 47% of students received scores below grade level (Florida Department of Education, n.d.). Similar to assessment results described for students in Minnesota, significant score differences can be seen among subgroups. Many other states report high percentages of students reading below grade level, including 53% in California (Department of Education State of California, n.d.), 41% in Nebraska (Nebraska Department of Education, n.d.), and 54% in New York State (New York State Education Department, 2024). These patterns can be seen across states of varying sizes, geographic regions, and demographic characteristics, leading to the enactment of dyslexia legislation in all fifty states (National Center on Improving Literacy [NCIL], n.d.). Additionally, many states have adopted Multi-Tiered Systems of Support (MTSS) to address and prevent wide-spread academic difficulties. An example of such efforts can be found in Washington State.

Efforts to Prevent and Remediate Reading Difficulties in Washington State

Washington State's MTSS framework includes five components: team-driven shared leadership; family, student, and community engagement; a continuum of supports; evidence-based practice; and data-based decision making (Flores et al., 2020). Washington's MTSS framework emphasizes the importance of using data to identify students who may benefit from additional supports, allocate resources, and inform instruction (Flores et al., 2020). This process is often referred to as universal screening.

What is Universal Screening for Reading?

Universal screening of reading is an increasingly common practice within schools (Silverman et al., 2021), and a significant body of research has documented its importance within a multi-tiered system of supports (MTSS) framework (Compton et al., 2006; Jenkins et al., 2007; Ochs et al., 2020; Rice et al., 2024; Vaughn et al., 2008). In universal screening, all students within a school participate in brief assessments of key academic skills (Missall et al., 2021; Shapiro & Clemens, 2023). Typically administered three times a year, universal screening measures for reading are used to identify students that are at-risk for learning difficulties (Ardoin & Christ, 2008; Jenkins et al., 2007; Kilgus et al., 2014; Silverman et al., 2021).

How does Universal Screening Address Limitations of NAEP?

Universal screening procedures address several limitations of NAEP. First, while NAEP reading tasks tend to measure background knowledge and vocabulary (NAGB, 2022) universal screening measures assess a wider variety of skills that contribute to reading proficiency, such as phonological awareness, phonics, fluency, and vocabulary (Ardoin & Christ, 2008; Good et al., 2001; Missall et al., 2021). Because universal screening measures of reading tend to assess a

wider variety of key reading skills, their results are better suited to inform instruction and intervention than results obtained by NAEP.

Additionally, NAEP scores are not reported at the school, classroom, or the student level, further limiting their instructional utility. In contrast, universal screening results can be examined at the school, grade-level, classroom, and student level (Deno et al., 2009), providing valuable information about the health of core instruction, where to allocate resources, and which students would benefit from additional supports. Quality universal screening measures include standardized administration and scoring guidelines to ensure that valid interpretations can be drawn from the assessment scores (Missall et al., 2021). There are many published universal screening measures available, with varying administration guidelines and score report formats. However, the usefulness of a universal screening measure, depends on its psychometric quality and the types of decision that are being made.

Chapter 2: Literature Review

Determining the Psychometric Quality of Universal Screeners of Reading

A core tenet of universal screening is the use of psychometrically sound assessment measures. Two important metrics for evaluating the quality of a universal screening measure include criterion validity and diagnostic accuracy (Glover & Albers, 2007; Jenkins et al., 2007; Kilgus et al., 2014; Outhwaite et al., 2024). Criterion validity is the extent to which a screener is correlated with another measure that is purported to measure the same construct adequately (Kilgus et al., 2014). Correlational evidence alone, however, is insufficient. Universal screening tools must also provide useful information for allocating instructional resources (Jenkins et al., 2007; Kent et al., 2019). As such, diagnostic accuracy is a crucial feature of universal screening measures (Kilgus et al., 2014). Diagnostic accuracy, also referred to as classification accuracy, is defined as a test's ability to correctly determine whether an individual is at-risk or not at-risk of not reaching a specific learning outcome, such as not passing a high stakes assessment (January & Klingbeil, 2020; Kent et al., 2019).

A major goal of assessments used for universal screening is to reduce decision-making error (Burns et al., 2023). Optimizing decision accuracy involves balancing sensitivity, specificity, positive predictive power, and negative predictive power (Burns et al., 2023; Kent et al., 2019; Kilgus et al., 2014; Ochs et al., 2020; Silverman et al., 2021). These four categories of decision accuracy are considered conditional probability statistics (VanMeveren et al., 2020), with values ranging from 0 to 1 (Kilgus et al., 2014).

Sensitivity refers to an assessments ability to identify students that are truly at-risk for reading challenges, while specificity relates to an assessments ability to identify students who are not truly at-risk for challenges in reading (Burns et al., 2023; Kilgus et al., 2014; Silverman et

al., 2021). Positive and negative predictive value relates to the proportion of students correctly identify as either at-risk and not at-risk, respectively (Glover & Albers, 2007; Kilgus et al., 2014). Consensus regarding adequate sensitivity and specificity levels has not yet been reached (January & Klingbeil, 2020; Kilgus et al., 2014); however, researchers have argued that, at minimum, a test's specificity value should be .70 (Kent et al., 2019) and sensitivity should be .80 or higher (National Center on Intensive Intervention [NCII], 2021).

Low sensitivity leads to the under-identification of students with reading difficulties, and thus prevents access to needed supports for these students (Catts et al., 2009). As such, researchers have argued that sensitivity is a particularly important aspect of universal screening (Burns et al., 2023). However, low specificity is also associated with negative outcomes. For example, providing interventions for students that do not truly need them may drain financial, logistical, and time related resources (Ford et al., 2017; Kent et al., 2019; Kilgus et al., 2014; Van Norman et al., 2017).

Finding efficient and effective ways to identify students with reading difficulties is a primary concern for educators (Burns et al., 2023). In addition to identifying students in need of supplemental instruction, universal screening measures can be used to predict performance on state assessments (Kent et al., 2019; Ochs et al., 2020; Van Norman et al., 2017) and assess the effectiveness of universal instruction (Missall et al., 2021). Because schools are often tasked with using finite resources to address a multitude of needs, effective universal screening procedures can help schools decide how and where to allocate limited resources (Kent et al., 2019; Ochs et al., 2020).

Data-Based Decision Making Using Universal Screeners of Reading

Data-based decision making (DBDM) is a process of collecting, interpreting, and using data to inform instruction (Jung et al., 2018; Majeika et al., 2024). Common types of decisions that are informed by data include which students should receive additional supports; how a student's skills are responding to intervention; and when interventions should be maintained, intensified, discontinued, or modified (Anderson & Borgmeier, 2010; Deno, 1985; Van Norman et al., 2018). DBDM can be found in a variety of effective instructional techniques (Jung et al., 2018), such as data-based program modification (Deno & Mirkin, 1977), experimental teaching (Casey et al., 1988), and precision teaching (Bruhn et al., 2020; Weisenburgh-Snyder et al., 2015). Typically, DBDM involves collecting student-level data, using a decision rule to evaluate progress toward a goal, and adapting instruction based on the data collected (Filderman et al., 2018; Fuchs et al., 1994; Hoogland et al., 2016). Further, the type and frequency of data collection methods may differ as a function of the type of decision being made. For example, when assessing a student's reading skills, triannual screening data may be sufficient for identifying students in need of supplemental supports, but more frequent data may be more appropriate for determining eligibility for special education (Ardoin & Christ, 2008).

Norm-referenced scores can support decisions related to resource allocation (Ardoin & Christ, 2008). Typically, students scoring below the 15th percentile are candidates for receiving small-group reading interventions (Ardoin & Christ, 2008). The use of decision rules has been associated with more effective instructional practices (Deno et al., 2001). DBDM using universal screening involves collecting school-wide data from all students (Bruhn et al., 2020; Missall et al., 2021). When measuring academic skills, the collection of curriculum-based measure (CBM) data at the class level has contributed to both effective instructional decision making and skill

growth among students (Stecker et al., 2005). Some evidence suggests that screening processes that utilize multiple reading measures may have higher classification accuracy than those that only use one measure (Jenkins et al., 2007). Although, using multiple measures requires additional time and financial resources (Barrett et al., 2024) and may not improve classification accuracy (VanDerHeyden et al., 2018).

Learning Assistance Program (LAP) Services

One of the purposes of universal screening is to identify students that would benefit from supplemental interventions, often referred to as tier 2 supports. In Washington state, these types of interventions can be provided through a Learning Assistance Program (LAP) service (OSPI, n.d.), a state-funded program provided to students who have not yet reached benchmarks in reading, writing, or math (OSPI, 2023, 2024). School districts use LAP funds to identify students that are eligible for supplemental services, determine the extent and cause of the academic need, and match this need with an evidence-based intervention (OSPI, 2024).

Washington state has developed an Integrated Student Support (ISS) protocol, which is meant to help districts coordinate LAP services across home, community, and school (OSPI, 2023, 2024). Washington requires that this protocol includes a needs assessment, description of community partnerships, and data regular data collection. The data collected are then reported to the state at the student, school, and district level (OSPI, 2023). The state allows the use of CBMs or computer adaptive tests (CATs) to calculate the academic growth of students participating in LAP services (OSPI, n.d.).

Universal Screening of Reading in Grades 4 and 5

When engaging in universal screening of reading for students in grades 3 through 5, two important considerations should be made. The first consideration is that core instruction in

reading focuses on understanding complex text, rather than learning foundational reading skills (Ochs et al., 2020); further, as students complete third grade, educators, administrators, and families must consider performance on state assessments (Kent et al., 2019; Van Norman et al., 2017). While a substantial body of research has investigated the importance of universal screening of reading skills in early elementary, less research has focused on universal screening considerations for upper elementary students (Kent et al., 2019) and the types of reading difficulties that emerge after grade 3 (Leach et al., 2003). However, the upper elementary years may be a particularly important period to engage in universal screening (Kent et al., 2019), as many students may first begin to demonstrate reading difficulties during the upper elementary years (Badian, 1999; Leach et al., 2003).

Reading Skills Assessed during Universal Screening

Universal screening measures should target skills that are addressed during classroom instruction (Clemens et al., 2015) and are predictive of later reading success (Jenkins et al., 2007). Universal screening of reading in early elementary grades tends to focus on phonological awareness, letter-sound knowledge, word reading fluency (Silverman et al., 2021), and vocabulary (Jenkins et al., 2007). Older students may experience reading challenges in the foundational skills such as phonics and word reading fluency, as well as vocabulary and comprehension (Leach et al., 2003; Vaughn et al., 2008).

For example, Leach and colleagues (2003) found that students that first demonstrated reading difficulties in grades 4 and 5 could be categorized into three groups: word reading difficulties, comprehension difficulties, and both word reading and comprehension difficulties. Specifically, the percentage of the sample that met criteria for each category was 35%, 32%, and 32%, respectively. The researchers also found evidence that these challenges emerged in later

grades and may not have been the result of misidentification in earlier grades (Leach et al., 2003). Thus, both foundational and higher-level reading skills may be important areas of screening for students in grades 3 through 5. Some screening measures of reading that can be administered in grades 3 through 5 include measures in the areas of word recognition, oral reading fluency, vocabulary, and comprehension (Curriculum Associates, 2022; University of Oregon, Center on Teaching and Learning, 2018).

Word Recognition

Decoding involves identifying individual sounds within words, matching these sounds with printed representations, and blending them together (Bell & Wheldall, 2022; Ehri, 2005; Gough & Tunmer, 1986). These skills continue to hold importance during the upper elementary years (Tortorelli et al., 2024). Word reading difficulties can impact both reading fluency and comprehension for upper elementary school students (Sabatini et al., 2019; White et al., 2021). As a beginning reader encounters a word for the first time, they may use decoding to sound out and read the word. When an individual matches sound to print, the word is more easily stored into their long-term memory, enabling them to read the word automatically, or by sight, when they encounter it in the future (Ehri, 2005). Skilled readers no longer need to sound out these words, but they can recognize them as single units (Ehri & Saltmarsh, 1995). Automaticity of word recognition may be facilitated by repeated exposures to a word (Ehri, 2005; LaBerge & Samuels, 1974), as well as having a strong understanding of its sound-spelling correspondence (Ehri, 2005). Ehri (2005) describes four phases of a word reading development: pre-alphabetic, partial alphabetic, full alphabetic, and consolidated alphabetic.

During the pre-alphabetic phase, children attempt to read words by remembering visual cues (Ehri, 2005). For example, in this phase a child might recognize common words printed

road signs, food and drink labels, or signs for restaurants not by matching sound to print, but by identifying salient visual cues (Mason, 1980). Children within the partial alphabetic phase may learn the names and sounds of letters; however, they have not yet developed full connections between sound and print (Ehri, 2005). In this phase, children may recognize the first and ending sound and printed letters (Savage et al., 2001), but still find reading vowel sounds within words challenging (Ehri, 2005).

When complete connections between letters and sounds are formed, children enter the full alphabetic phase. Children in this phase can segment sounds within words and match them to print, which allows them to store these letter-sound relationships in their memory (Ehri, 2005). This ability is an essential component of word reading (Ehri, 2005; Liberman et al., 1974; Share et al., 1984). Thus, the full alphabetic phase is marked by an increase in word reading accuracy and the ability to differentiate similarly spelled words (Ehri, 2005; Ehri & Wilce, 1987).

As word recognition skills become more advanced, readers transition from processing individual letter-sound relationships to recognizing letter patterns (Bhattacharya & Ehri, 2004), which is likely to aid longer, multi-syllabic word reading (Ehri, 2005). The recognition of larger units of print is a characteristic of the consolidated alphabetic phase (Ehri, 2005; Kuhn & Stahl, 2003). These larger units include syllables, onset and rimes, and the meaning units within words, referred to as morphemes (Ehri, 2005). The application of word reading skills to multisyllabic words is particularly important as students enter upper elementary grades. For example, students in 3rd grade encounter significantly more multisyllabic words compared to students in earlier grades (Kearns & Hiebert, 2022). The immediate recognition of groupings of letters, also referred to as orthographic awareness, also contributes to text reading fluency (Kim, 2015). As such, as children enter this phase, the focus of reading may once again shift from accurate

reading to fluent reading (Kuhn & Stahl, 2003). While word reading is an important component of overall reading achievement, for upper elementary students, word reading measures may have limited utility related to state assessments (Kent et al., 2019).

Oral Reading Fluency

Oral reading fluency can be characterized by the automaticity and prosody with which an individual converts print to speech (Fuchs et al., 2001; Kuhn & Stahl, 2003; LaBerge & Samuels, 1974; Perdomi et al., 2024; Schwanenflugel et al., 2006). A common definition of reading fluency includes three main components: accuracy, rate, and expression (Dowhower, 1991; Kuhn & Stahl, 2003; Maki & Hammerschmidt-Snidarich, 2022; Shinn et al., 1992). Typically, accuracy is measured by the percentage of words read correctly (Daane et al., 2005). Rate refers to the number of words read over a specific period of time, which often one minute (Daane et al., 2005). Reading fluency is often determined by calculating the number of words read correctly per minute (Hudson et al., 2009; Jenkins et al., 2003; Washburn, 2022), although several researchers have advocated for the consideration of prosody in assessing reading fluency. Prosody refers to the use of intonation, pauses, and stress when reading, characteristics that help children organize texts into units of meaning (Allington, 1983; Daane et al., 2005; Dowhower, 1991; Kuhn & Stahl, 2003; Shhub et al., 2023). As a child's reading becomes more prosodic, their reading begins to reflect the emphasis, phrasing, and inflection of oral speech (Fuchs et al., 2001; Kuhn & Stahl, 2003). Rating scales are a common tool for assessing prosody (Wolters et al., 2022).

Additionally, researchers have made a distinction between oral passage reading fluency (OPR) and oral reading fluency (ORF; Fuchs et al., 2001; Jenkins et al., 2003; Kim, 2015; Kim & Wagner, 2015; Kim et al., 2014). OPR relates to reading connected text, while ORF relates to

reading words in isolation (Altani et al., 2020; Jenkins et al., 2003; Kim, 2015; Kim & Wagner, 2015). Although the two tasks overlap (Altani et al., 2020; Jenkins et al., 2003), ORF may be more salient to reading comprehension for students in early elementary grades, while OPR may play a mediating role in comprehension by the time for older students (Jenkins et al., 2003; Kim & Wagner, 2015; Schwanenflugel et al., 2006). Several studies suggest that ORF may be correlated with rapid naming tasks (Altani et al., 2017; de Jong, 2011; Georgiu et al., 2013). Jenkins and colleagues (2003) found that OPR was a better predictor of reading comprehension than isolated word-reading for grade 4 grade students. Additionally, OPR, through its relation to decoding, may be a distinct construct from reading comprehension by fifth grade (Shinn et al., 1992). In fact, OPR may act as a bridge between ORF, listening comprehension, and reading comprehension in grades 3 and 4 (Kim & Wagner, 2015). Stanovich and colleagues (1984) found that words are recognized faster when presented in context. Fluent reading has also been associated with improved comprehension (Daane et al., 2005), with researchers suggesting that accurate and automatic word reading allows children to direct cognitive resources to reading for understanding (Fuchs et al., 2001; Kim, 2015; LaBerge & Samuels, 1974; Schwanenflugel et al., 2006).

From an assessment perspective, Fuchs and colleagues (1988) found that measures of oral reading fluency had higher correlations with criterion comprehension measures compared to informal comprehension tasks, such as story recall, cloze, and question answering procedures. In addition to its association with reading comprehension (Fuchs et al., 1988), ORF has been described as a helpful metric for assessing reading achievement due to its ability to compare performance across students, monitor performance of individual students, and guide educational decision making (Fuchs et al., 2001). Further, the most rapid growth in ORF appears to occur in

early elementary school, with some studies finding less growth occurring by grade 4 (Fuchs et al., 1993). When used for universal reading screening, some research indicates that reading fluency measures alone have limited classification accuracy, with specificity values ranging from .31 to .54 in some studies (Kent et al., 2019).

Vocabulary

Vocabulary, or the knowledge of word meanings, is an integral component of reading and overall academic achievement (Beck et al., 2013; Biemiller & Boote, 2006; Blachowicz et al., 2006; Kuder, 2017; National Institute of Child Health and Human Development (NICHD, 2000). Beck and colleagues (2013) suggested that vocabulary words can be categorized into three tiers. Tier 1 words are those that are used frequently in conversation, requiring little formal instruction to learn; tier two words are foundational for using and understanding language, occur across domains, and are more likely to appear in written form than through conversation; tier three words can be considered domain specific vocabulary words and occur less frequently than words at tier one and tier two. Tier 2 words are optimal candidates for vocabulary instruction (Beck et al., 2013).

Effective vocabulary instruction involves the direct teaching of word meanings using concise definitions, as well as varied activities that deepen knowledge of specific words, such as repeated exposure to words, word association, sentence generation, and definition matching tasks (Beck et al., 1982, 2013; Blachowicz et al., 2006; McKeown et al., 1985; NICHD, 2000; Sanchez & O'Connor, 2022). In addition to direct instruction in vocabulary words, students may benefit from learning strategies for finding the meaning of unknown words (Graves et al., 2018). While vocabulary can also be acquired incidentally through naturally occurring contexts, such as reading a wide range of texts and exposure to oral language, these contexts are not always

optimal for learning new vocabulary words sufficiently (Beck et al., 2013; Nagy et al., 1985). The type of context where words are likely to occur shifts as students become older. For example, younger learners may learn vocabulary through oral language interactions. However, for older students, incidental word learning is more likely to occur through text (Beck et al., 2013).

Both the depth and breadth of word knowledge play important roles in vocabulary development (Coyne et al., 2009; Lee, 2024; Li et al., 2021). Depth refers to how well a word's meaning is understood (Coyne et al., 2009). Attention to the depth of word knowledge can aid reading comprehension (Lee, 2024). However, focusing on depth can be time consuming, and thus instruction may only be feasible for a limited number of words (Lee, 2024). Breadth refers to the number of words an individual knows (Coyne et al., 2009; Li et al., 2021). Instruction that focuses on breadth may incorporate word learning in the context of read aloud activities (Biemiller & Boote, 2006; Coyne et al., 2009). This type of instruction uses brief definitions to explain words within a story.

Other oral language skills, such as syntax and listening comprehension, may also serve as important predictors of later reading achievement (Silverman et al., 2021). Listening comprehension, for example, has been identified as a predictor of text reading fluency (Kim, 2015). Li and colleagues (2021) found that vocabulary depth directly predicted reading comprehension, while breadth influenced listening comprehension which in turn influenced reading comprehension. Silverman and colleagues (2021) found that oral language tasks like sentence repetition predicted reading comprehension in grades 1 and 3. Of the many benefits of vocabulary and oral language skills, their relationship to comprehension (Ahmed et al., 2016;

NICHD, 2000) has been identified as one of its most important features (Beck et al., 1982; Biemiller & Boote, 2006; Quinn et al., 2020).

Comprehension

Reading comprehension refers to the ability for an individual to create mental representations or meaning from information found in a text (Hattan et al., 2023; Hoover & Tunmer, 2018; NICHD, 2000). Considered by many to be the primary goal of reading instruction, reading comprehension is complex, subsuming a variety of cognitive, academic and linguistic subskills (Cromely et al., 2010; Elleman & Oslund, 2019; NICHD, 2000). Another component of reading comprehension is the active monitoring of understanding as one interacts with text, a process that can be facilitated through cognitive strategy instruction (NICHD, 2000). For older students, reading comprehension becomes an increasingly important predictor of reading success (Cho & Barrett, 2024; Jenkins et al., 2007).

A variety of models and theories describe how reading comprehension develops (Ahmed et al., 2022). The Simple View of Reading (SVR; Gough & Tunmer, 1986; Hoover & Gough, 1990), for example, stipulates that reading comprehension is a combination of word reading and language comprehension abilities (Ahmed et al., 2016; Duke & Cartwright, 2021; Gough & Tunmer, 1986; Hoover & Gough, 1990; Kelso et al., 2022). Some researchers argue that comprehension is facilitated by two processes: construction and integration (Hattan et al., 2023; Kintsch, 1988). Construction refers to the meaning that a reader derives from a specific text, while integration involves connecting that meaning to prior knowledge (Hattan et al., 2023; Kintsch, 1988). The construction-integration model of reading, also referred as text processing, has been used to develop comprehension interventions such as the Promoting Adolescents' Comprehension of Text (PACT) program (Vaughn & Wanzek, 2024).

The direct and inferential mediation (DIME) model of reading comprehension (Cromely et al., 2010) expands on the Simple View of Reading and text processing theories by arguing that reading comprehension involves background knowledge, inferencing, and reading strategies alongside word reading and language comprehension skills (Ahmed et al., 2016, 2022; Cromely & Azevedo, 2007; Cromely et al., 2010; Elleman & Oslund, 2019). According to the DIME model, these subskills are interrelated and contribute to reading comprehension at varying degrees depending on the age of the individual (Ahmed et al., 2016; Bell & Wheldall, 2022).

Types of Universal Screening Measures of Reading

Curriculum-Based Measures

Curriculum-Based Measures (CBMs) are standardized, repeatable, and brief measures that assess discrete academic skills, allowing for educators to assess a student's current level of reading proficiency, as well as growth over time (Deno, 2003; Deno et al., 2001; Fuchs & Fuchs, 1992; Hosp et al., 2016; Jung et al., 2018; Missall et al., 2019; Wayman et al., 2007). CBMs feature standardized directions, scoring rules, and interpretation guidelines that aid in instructional decision making (Hosp et al., 2016; Van Norman et al., 2018). The methods used to construct CBMs have been described as a combination of traditional psychometric approaches, involving an emphasis on reliability and validity, with behavioral assessment approaches that include visual analysis of data, repeatability, and data collected over time (Deno et al., 2001; Fuchs & Fuchs, 1992; Hosp et al., 2016; Jenkins & Fuchs, 2012). CBMs were developed to be time-efficient, practical, inexpensive, and easy to use for monitoring the effectiveness of instruction (Deno et al., 1982; Jenkins & Fuchs, 2012). Decades of studies have demonstrated evidence of the technical adequacy, utility, and feasibility of CBMs (Deno et al., 2001; Hosp et al., 2016; Jenkins & Fuchs, 2012). For example, researchers have found evidence of criterion

validity with commercial reading assessment (Fuchs et al., 1988), construct validity (Shinn et al., 1992; Wayman et al., 2007), and discriminant validity (Deno et al., 1982; Shinn et al., 1987).

CBMs used to measure reading performance, CBM-R, are often timed and typically involve calculating the number of words read correctly by a student per minute (Deno et al., 1982, 2001; Hosp et al., 2016; Van Norman & Forcht, 2024), a useful metric for measuring response to intervention (Van Norman et al., 2018). Additionally, a useful characteristic of CBM-R is that it mirrors the types of activities that students may be asked to do during classroom instruction, leading to the collection of instructionally relevant data (Hosp et al., 2016).

When used for universal screening purposes, a set of three CBM-R probes are often administered and a student's median score is used for interpretation (Ardoin & Christ, 2008); although, some researchers argue that the use of multiple probes in one screening is unnecessary to estimate the level of a student's reading performance (Ardoin & Christ, 2008). CBM-R results for a student can be interpreted in comparison to their peers, an established standard of achievement, or to their prior CBM-R results (Ardoin & Christ, 2008; Deno et al., 2001; Hosp et al., 2016). Additionally, CBM-R universal screening data can be used to determine both a students' level of reading achievement and rate of growth over time in reading (Ardoin & Christ, 2008). The use of multiple probes requires that CBM-Rs show evidence of alternate form reliability (Deno et al., 2001; Fuchs, 2017; Wayman et al., 2007), which assumes that the various passages being used for screening are relatively equivalent in terms of difficulty, mean student performance, and the standard deviation of student scores (Ardoin & Christ, 2008; Betts et al., 2009; Wayman et al., 2007). CBMs lend themselves to visual analysis through graphing, which

is increasingly important for understanding how skills respond to intervention (Oslund et al., 2021).

CBM-R and State Assessments

Predicting performance on state assessments is often a primary focus for educators (Decker et al., 2014; Hintze & Silberglett, 2005; Neese, 2022). Many studies have examined the predictive utility of CBM-R in relation to performance on year-end state assessments (Kilgus et al., 2014; Neese, 2022; Shin & McMaster, 2019). Good and colleagues (2001) found that 3rd graders who read at least 110 words correct per minute (WCPM) on a CBM were likely to meet grade level expectations on the state assessment, while 3rd graders who read 70 WCPM were less likely to score in the expected range on the state assessment. Hintze and Silberglett (2005) examined the predictive utility of CBM-R using three statistical methods: discriminant analysis, logistic regression, and receiver operator characteristics (ROC) curve analysis, finding that CBM-R scores predicted performance on state assessments regardless of the statistical method used. Moderate to strong correlations between CBM-R and state assessments were found for students in grades 3 and 5 (Shapiro et al., 2006), students in grade 4 (McGlinchey & Hixson, 2004; Stage & Jacobsen, 2001), and similar findings were observed for students in grade 7 (Hunley et al., 2013).

CBM-R has been found to be a consistent predictor of state test performance (Yeo, 2010) for students in elementary (Good et al., 2001; Hintze & Silberglett, 2005; Missall et al., 2019), middle (Decker et al., 2014; Espin et al., 2010; Hunley et al., 2013; Shin & McMaster, 2019), and high school (Shin & McMaster, 2019). Research suggests, however, that the prediction accuracy of CBM-R can vary as a function of test publisher and whether criterion- or norm-referenced cut scores are used (Ford et al., 2017). Other factors may impact the predictive utility

of CBM-R, such as sample size of the study, the proportion of ML students included, and the time interval between the CBM-R and state assessment administration (Yeo, 2010). More recent research has examined the predictive utility of computerized applications of CBM-R, finding that both traditional and computer-based forms of CBM-R had high area under the curve (AUC) values, but low sensitivity (Nese, 2022).

Computer Adaptive Testing

Increasingly, schools are administering Computer Adaptive Tests (CATs) as universal screeners of reading (Clemens et al., 2015; Sutter et al., 2020). Students taking CATs will see different items of varying difficulty depending on how they responded to previous items (Clemens et al., 2015; Meijer & Nering, 1999; Shapiro et al., 2015; Van Norman & Forcht, 2024). For example, if a student answers test items incorrectly, they will be presented with less challenging items; however, the difficulty of the items increases when a student answers test questions correctly (Luo et al., 2017; Ochs et al., 2020; Shapiro et al., 2015; Sutter et al., 2020). This process may allow for more precise estimation of a student's reading proficiency.

CATs are developed using Item Response Theory (Meijer & Nering, 1999; Shapiro et al., 2015). IRT is used to help predict an examinee's response to a specific item (Meijer & Nering, 1999). Learning progressions, or the hierarchy of skills necessary to obtain proficiency in a domain, are used to develop items of varying difficulty levels on CATs (Shapiro et al., 2015). To accommodate the large number of potential item presentations, CATs often have a large item pool from which to draw (Shapiro et al., 2015).

Additional benefits of CATs relate to feasibility and standardization. Because many CATs can be administered and scored by the automated program itself, these measures may limit opportunities for administration errors, as well as allow educators devote more time to planning

and implementing effective instruction (Clemens et al., 2015). Further, the efficient calibration of item difficulty may lead to shorter tests, helps identify a student's strengths and needs, and enhances measurement precision (Clemens et al., 2015; Meijer & Nering, 1999; Shapiro et al., 2015)

These characteristics may support instruction, limit student frustration, and facilitate effective data interpretation. Some researchers argue that CATs allow for the assessment of a wide variety of reading skills, helping educators choose and implement more targeted reading interventions (Ochs et al., 2020). CATs can also be used to make comparisons across grades (Shapiro et al., 2015) and have been shown to be moderately predictive of end-of-year and state exams (Clemens et al., 2015; Ochs et al., 2020); however, CATs may be more costly, less predictive, and take longer to administer than paper-based CBMs (Clemens et al., 2015; Meijer & Nering, 1999).

State Assessments

State assessment results from previous years represent another avenue for universal screening of reading skills (Denton et al., 2011; VanMeveren et al., 2020; Van Norman et al., 2017). As students progress through grade levels, information from the previous years' state assessment results could be used to identify those in need of additional supports and predict performance on the next state assessment (Paly et al., 2022; VanMeveren et al., 2020; Vaughn et al., 2008), although most of these tests were not designed for such purposes (Nelson et al., 2017). The use of historical data, such as state assessment results, could decrease the amount of time, money, and logistical resources needed for universal screening (Paly et al., 2022; VanDerHeyden et al., 2017; VanMeveren et al., 2020). These resources could then be directed toward

instruction. Additionally, the predictive utility of prior state testing results may be enhanced by using locally constructed cut scores (Van Norman et al., 2017).

Several studies have investigated the utility of prior-year's state assessment results in predicting performance on the next state assessment (Nelson et al., 2017). Paly and colleagues (2022) found that using prior year state assessment results both alone and combined with a CBM resulted in acceptable levels of specificity; however, overall classification accuracy was highest when using state assessment results as the first step in a multiple-gating process, similar to findings from other studies (Van Norman et al., 2017). The authors of the study acknowledged the potential usefulness of state assessment results for universal screening in upper elementary and beyond, but stated that their results may not generalize to other state tests (Paly et al., 2022).

Van Norman and colleagues (2017) compared prior state test results to a CAT in predicting performance on the next state assessment. Their study found that the two measures had similar values of sensitivity and specificity (Van Norman et al., 2017). Another study found that prior state test results yielded acceptable levels of sensitivity and specificity when using researcher-developed cut scores; these cut scores resulted in acceptable classification accuracy when applied to future cohorts of students (Nelson et al., 2017). One study compared measures of oral reading fluency, silent reading fluency, and verbal knowledge to prior state assessment results when used for universal screening, finding that previous state test performance was the best predictor of future state test performance (Denton et al., 2011). Further, the addition of other reading measures explained only 5% of the variance in later state test scores (Denton et al., 2011). While evidence suggest that previous year's state test results can effectively predict performance on the next state test, methods of comparing stand-alone universal reading screeners

with historical state testing data should be replicated with other commonly used measures (Paly et al., 2022; Van Norman et al., 2017), as well as other state tests (Denton et al., 2011).

Commonly Used Measures for Universal Screening of Reading

iReady Reading Diagnostic

The iReady Reading Diagnostic assessment is a CAT that measures a student's skills in a variety of reading domains, including phonological awareness, phonics, high frequency words, vocabulary, comprehension of informational texts, and comprehension of narrative texts (Curriculum Associates, n.d.c.). The iReady Reading Diagnostic provides an overall estimate of a student's instructional level in reading, as well as grade level placements for each domain (Curriculum Associates, n.d.c.) According to the test publisher, iReady can be used for screening, progress monitoring, and as an accountability measure (Curriculum Associates, 2021). Results can be viewed at the district, school, class, and individual student level. Score reports include recommendations for instructional groupings and lesson planning resources, and information to share with families (Curriculum Associates, n.d.c., 2021). iReady results can be interpreted using both criterion referenced and norm-referenced scores, and used to select appropriate texts, monitor growth toward grade-level proficiency, and predict later academic performance (Curriculum Associates, n.d., n.d.c., 2020, 2021). iReady Reading Diagnostic norms are reviewed every year to determine if updates are needed (Curriculum Associates, 2023).

A series of studies to demonstrate evidence of iReady's predictive utility in relation to state assessments. One such study examined the correlation between the iReady Reading Diagnostic and the Colorado Measure of Academic Success (CMAS; Colorado Department of Education, 2021). This study found spring correlations of $r = .86$ at grade 3, $r = .86$ at grade 4,

and $r = .82$ at grade 5 (Curriculum Associates, 2022d). Similar studies have been conducted in Mississippi, correlations ranging from $r = .78$ to $.83$ in grades 3 through 5 (Curriculum Associates, 2019); New Hampshire, correlations ranging from $r = .81$ to $.83$ in grades 3 through 5 (Curriculum Associates, 2024); and in Hawai'i, correlations reported as $r = .83$ for all three grades (Curriculum Associates, 2022c).

Dynamic Indicators of Basic Early Literacy Skills, 8th Edition Oral Reading Fluency (DIBELS-ORF)

The Dynamic Indicators of Basic Early Literacy Skills, 8th Edition (DIBELS-8) are a set of CBMs designed to measure foundational literacy skills (University of Oregon, n.d., 2018). DIBELS emerged out of the work of Kaminski and Good (1996), and they are used widely as universal screeners of reading (Munger & Murray, 2017). DIBELS-8 measures can be administered to students in grades kindergarten through 8 (University of Oregon, n.d.). Consistent with research on CBMs, these measures are meant to be easy to administer, sensitive to growth, and repeatable (Kaminski & Good, 1996). The psychometric quality of DIBELS-8 has been described in terms of test-retest reliability, alternate-form reliability, predictive validity, and classification accuracy (University of Oregon, 2018). These metrics will be described in detail in the Method section. One of the most widely used CBMs is oral reading fluency (ORF; Nese & Kamata, 2021).

Test-retest reliability for DIBELS-ORF was reported as $r = .94$ for grade 3, $r = .91$ for grade 4, and $r = .91$ for grade 5 (University of Oregon, 2018). Alternate form reliability coefficients were reported as $r = .95$ for grades 3 through 5. Correlations between DIBELS-ORF and the ORF subtest from previous versions of DIBELS were $r = .92$ for grade 3, $r = .90$ for grade 4, and $r = .95$ for grade 5 at the beginning of the school year (University of Oregon, 2020).

Present Study

Educators should ensure that universal reading screening measures are both practical and technically adequate (Glover & Albers, 2007; Missall et al., 2021). Factors that relate to practicality include the time it takes to train staff members, assessment administration, and the financial cost of these measures (Paly et al., 2022). Given the popularity of DIBELS-ORF and iReady Reading Diagnostic as universal screeners of reading, determining which measure is a better predictor of end-of-year state performance will have important technical and practical implications for teachers, administrators, and students.

For example, iReady costs approximately \$7.25 per student, can take up to 45 minutes to administer, and requires access to computer and internet (NCII, n.d.), while DIBELS-ORF is freely available, can be completed in one minute, and can be administered with paper and pencil materials (University of Oregon, n.d., n.d.b., 2023). If DIBELS-ORF proved to be a better predictor of state assessment performance, then switching from using iReady to DIBELS-ORF would allow students to spend more time receiving instruction, while sparing financial, time, and technological resources that could be utilized for other tasks. However, DIBELS-ORF requires one-to-one administration, additional time for scoring, and may be prone to passage effects (Cummings et al., 2013; Petscher & Kim, 2011; NCII, n.d.b.). iReady, in contrast, can be group administered, has automated assessment instructions, score reports can be provided automatically, and uses an algorithm to select appropriate test items for individual students (NCII, n.d.). If iReady proved to be a better predictor of state assessment performance, then switching from DIBELS-ORF to iReady would require less time spent engaging in assessment preparation, administration, and scoring for educators. In schools that use both measures for screening, understanding which one is a better predictor of state test performance may allow

them to choose the most effective one, which in turn would decrease time dedicated for screening while increasing time dedicated for instruction.

While these questions have clear educational implications, no independent, peer-reviewed studies have compared DIBELS-ORF and iReady measures in terms of predictive validity related to the SBA-ELA. iReady's test publisher conducted a study examining the reading diagnostic's predictive relationship to the SBA for students in nine school districts in Connecticut (Curriculum Associates, 2023b). The study found correlations between the fall iReady Reading Diagnostic and the SBA to be $r = .82$ for grades 3-4 and $.81$ for grade 5 (Curriculum Associates, 2023b). However, the test publishers did not report their procedures or statistical analysis for this study. Additionally, they did not compare iReady to any other screening measure. A similar study was conducted with 49 school districts across California, Delaware, Oregon, and Washington, finding correlations of $r = .82$ for grade 3, $r = .81$ for grade 4, and $r = .81$ for grade 5 (Curriculum Associates, 2022d). Consistent with the previous study, analyses, procedures, and a detailed description of their results were not reported.

Roehrig and colleagues (2008) examined the predictive utility of DIBELS-ORF related to a state assessment in Florida for students in grade 3. The authors found AUC indices ranging from $.82$ to $.84$ (Roehrig et al., 2008). Another study identified a median predictive validity coefficients of $.68$ (Goffreda & DiPerna, 2010). Studies conducted by the test publisher have resulted in similar findings, with predictive validity coefficients ranging from $r = .63$ to $.83$ for students in grades 3 and 5 (University of Oregon, 2018). However, none of the reviewed studies compared DIBELS-ORF and iReady Reading Diagnostic in terms of predictive validity.

Additionally, there is evidence that a state's prior year assessment can greatly predict performance the following year (Denton et al., 2011; Van Norman et al., 2017). However, many

state assessments were not designed for this purpose. As such, educators cannot assume that performance on a state assessment will be predictive of later performance on the same assessment if that measure has not been examined for that purpose. If the prior year's state assessment scores predict later performance just as well as measures constructed specifically for universal screening of reading, then financial, staffing, and time resources could be reallocated to support instruction. Lastly, schools should also ensure that screening measures accurately predict performance on assessments for students participating in LAP services, as accurately predicting the academic trajectory of these students can support DBDM – namely, when to start, intensify, or discontinue a reading intervention.

Research Questions

Based on the foregoing discussion around the practical use and predictive utility of universal reading screeners, this study aimed to evaluate the following research questions.

- 1) To what extent do three universal reading measures – prior year SBA-ELA, fall iReady Diagnostic Reading, and fall DIBELS-ORF – *directly* predict spring 2024 SBA-ELA scores for students in grades 4-5?
- 2) To what extent is the predictive utility of these reading measures different across grades levels as well as students identified for the district's Learning Assistance Program (LAP)? In other words, do grade and/or student LAP status moderate the relationships between the three indicators and spring 2024 SBA-ELA scores?
- 3) To what extent do the iReady Diagnostic Reading and DIBELS-ORF *uniquely* predict spring SBA-ELA scores, above and beyond prior year SBA-ELA scores, and are the differences in magnitudes among the unique reading predictor effects statistically significant?

Chapter 3: Method

Sample

This study is a secondary analysis of extant data from one suburban public school district in Washington state that had reached out to the university's educational psychology researchers to explore the district's K-5 elementary literacy instruction policies and practices. After IRB approval and a data sharing agreement was signed by both organizations, the district transferred assessment and demographic data to the university through a secure, electronic link. The district's adopted core literacy curricula was *Wonders* (McGraw Hill, 2023) and all assessments were administered during regular school scheduling by teachers.

According to state's Office of the Superintendent of Public Instruction (OSPI), during the 2023-2024 academic year (same year as this study's sample grade 4-5 data), 49% of all K-12 enrolled students were female, 24% were students of color, 3% received multilingual services, 15% received special education services, and 23% were from low-income households. The present study analyzed data for all $N = 1,191$ enrolled grade 4-5 students from 57 classrooms across 10 schools who had complete data available, representing 92% of the total number of students enrolled in third or fourth grade in 2023 or fourth and fifth grade in 2024. As shown in Table 1, for this study's sample, $n = 566$ ($n = 48\%$) were female, $n = 245$ (21%) were students of color, $n = 33$ (3%) received multilingual services, $n = 197$ (17%) received special education services, and $n = 122$ (10%) were identified for the Learning Assistance Program (LAP).

Measures

Smarter Balanced Assessment – English Language Arts (SBA-ELA)

The Smarter Balanced Assessment (SBA; Smarter Balanced Assessment Consortium, n.d.) is a computerized adaptive test (CAT) that measures performance in both math and English

Language Arts (ELA). All students in grades 3 through 10 in Washington state are mandated to complete the SBA annually in the spring (OSPI, n.d.b) to assess students' college and career readiness (Smarter Balanced Assessment Consortium, n.d.). The ELA portion includes items that measure reading, writing, listening, and research inquiry (Smarter Balanced Assessment Consortium, n.d.b.). For students in grades 4 and 5, the SBA-ELA takes approximately 3.5 hours to complete (Smarter Balanced Assessment Consortium, 2017).

Items assessing reading skills require students to read narrative and informational passages, and to answer multiple choice questions about each passage read such as identifying the passage's main idea, using passage details to make inferences, determining the meaning of words in the passage, and identifying the text's structure (Smarter Balanced Assessment Consortium, n.d.b.). The writing items require students to revise passages, use multiple sources to write a story or an essay, and finish an incomplete passage (Smarter Balanced Assessment Consortium, n.d.b.). Items assessing listening skills require students to listen to, answer questions about, and identify the main purpose of several oral presentations. Finally, research and inquiry items require a student to support claims using evidence, identify relevant information to include in a research report, and write an informational article using multiple sources (Smarter Balanced Assessment Consortium, n.d.b.).

In addition to scale scores, schools and parents are also provided with Achievement Level Descriptors (ALD) for their children that correspond to ranges of the scale, from Level 1 – Level 4 (Smarter Balanced Assessment Consortium, 2013). Students scoring within Level 3 or higher are considered as having met the state's expectations. Table 2 provides an overview of SBA-ELA reading scale score ranges, levels, and expected skills for students in grades 4-5.

Scoring. Students complete the SBA on a computer, with each student receiving a personalized set of items based on the difficulty and accuracy with which they answered previous items until the student's skill level is estimated with sufficient precision to stop. Scale scores are derived automatically via item response theory modeling.

Psychometric evidence. According to the SBA technical manual (Smarter Balanced Assessment Consortium, 2022), the SBA reports scale score reliabilities in terms of “marginal reliability”, an IRT-derived metric that describes how precise an assessment is (Hayley et al., 2010). Like Cronbach's alpha, marginal reliabilities range from 0 (pure error) to 1 (perfect measurement), with .70–.80 considered acceptable for low-stakes decisions, .80–.90 considered good for classroom or institutional decisions, and >.90 required for high-stakes decision-making (e.g., licensing, admissions). According to their technical manual, marginal reliabilities for the SBA-ELA scale scores were .91 and .92 for grades 4 and 5, respectively (Smarter Balanced Assessment Consortium, 2022). According to the Smarter Balanced Assessment Consortium (2022), the SBA-ELA demonstrates evidence of strong classification accuracy. For students in grade 4, classification accuracy values for the SBA-ELA were .99 for reading, .97 for writing, .97 for listening, .98 for research and inquiry, and .80 overall. For students in grade 5, classification accuracy values for the SBA-ELA were .97 for reading, .98 for writing, .97 for listening, .99 for research and inquiry, and .81 overall.

iReady Reading Diagnostic Assessment

The iReady Reading Diagnostic assessment is also a CAT that can be administered to students in kindergarten through grade 12 (Curriculum Associates, n.d.). It can be group administered and takes 30-60 minutes to complete, according to information provided by the publisher (Curriculum Associates, n.d.b.). Additionally, the publishers recommend that this

assessment be administered three times a year (Curriculum Associates, n.d.b.). Four to eight hours of training is required for sites administering the iReady Reading Diagnostic.

Subtests. The full set of subtests available for the reading domain include Phonological Awareness, Phonics, High Frequency Words, Vocabulary, Comprehension of Informational Text, and Comprehension of Literature (National Center on Intensive Intervention [NCII], n.d.). Students in grades 4 and 5 complete only the latter three subtests, starting with Vocabulary first, then Comprehension of Literature, and then Comprehension of Informational Text (Curriculum Associates, 2022). However, if a student receives an overall diagnostic scale score from these three subtests that is below 511 (a level considered as below third-grade skills), students then complete the Phonics subtest. If they receive a score on this subtest below 421 (below first-grade skills), the student then also completes the High Frequency Words subtest (Curriculum Associates, 2022).

The Phonological Awareness subtest requires students to recognize rhymes, blend and segment sounds, and manipulate phonemes. Audio directions are presented to the student, along with pictures (Curriculum Associates, 2022). While completing the Phonics subtest, a student listens to a word or phoneme provided by an audio recording and matches the sound with its visual representation. During the High Frequency Words subtests, students are asked to either spell or to select a frequently used word presented by an audio recording. The Vocabulary subtest requires a student to read a passage, determine the meaning of an underlined word, and select the definition of the word from several options. Finally, students read and answer multiple choice questions about narrative and informational passages while completing the comprehension subtests (Curriculum Associates, 2022).

Scoring. When the iReady Reading Diagnostic is completed, a student's performance is scored automatically. Student scores are color-coded to indicate the level of intervention support recommended for each student. For example, scores colored green indicate that the student's reading development is met with tier 1 instruction, yellow indicates that tier 2 supports are recommended, and scores that are red indicate that tier 3 supports are recommended (NCII, n.d.). In addition to this color-coded system, criterion-referenced grade-level scores are also provided to allow educators to see the alignment between student skills and the core curriculum. Finally, percentile ranks are included, allowing for normative comparisons (Curriculum Associates, n.d.; NCII, n.d.).

Psychometric evidence. According to information submitted to NCII by the publisher, the iReady Reading Diagnostic assessment has good supporting reliability and validity evidence (Curriculum Associates, 2020a, 2022a; NCII, n.d.). The median marginal IRT-based reliability was reported as .97 for grades 3 through 5, and test-retest reliability was reported as .86 for both grades 4 and 5. In terms of subgroups, split-half reliabilities for Asian students were .80 for grade 4 and .79 for grade 5; for Black students, split-half reliabilities were .77 for grade 4 and .78 for grade 5; and for Latine students, .76 for grade 4 and .79 for grade 5.

The iReady Reading Diagnostic has also shown a strong positive correlation with other assessments, including the SBA-ELA (e.g., $r = .81$ for grades 4-5) as well as with the Alabama Comprehensive Assessment Program (ACAP; Alabama State Department of Education, 2024), with $r = .84$ for grades 4-5 (Curriculum Associates, 2022b). Similarly, the iReady was correlated at $r = .82$ for grade 4 and .81 for grade 5 with the Massachusetts Comprehensive Assessment System (MCAS; Massachusetts Department of Elementary and Secondary Education, n.d.) (Curriculum Associates, 2022c).

Dynamic Indicators of Basic Early Literacy – Oral Reading Fluency (DIBELS-ORF)

The Dynamic Indicators of Basic Early Literacy, Eight Edition (DIBELS) is a collection of curriculum-based measures (CBMs) that assess reading skills such as phonemic awareness, phonics, fluency, and comprehension (University of Oregon, Center on Teaching and Learning, 2018). The Oral Reading Fluency (ORF) subtest is individually administered to students in grades 1 through 8 (University of Oregon, 2023b).

The DIBELS-ORF takes approximately only one minute to administer (University of Oregon, 2023b). The required materials for this assessment include a scoring book, student forms, writing utensil, clipboard, and timer. ORF, along with other DIBELS subtests, can be downloaded for free from the DIBELS website (University of Oregon, n.d.) and can be used for both universal screening and progress monitoring. When used for universal screening, only one ORF probe is required per benchmark period (University of Oregon, n.d.b). Additionally, DIBELS measures can be administered in paper or digital format (University of Oregon, n.d.b).

Scoring. While the student reads, the test administrator makes note of any errors the student makes. Errors are recorded if a student mispronounces a word, does not blend sounds within words, omits a word, or reads words in the wrong order (University of Oregon, 2023b). ORF produces two scores. One score is the number of words read correctly (wcpm), and the second score is a percentage derived from the number of words read correctly divided by the total number of words (University of Oregon, 2023b).

Psychometric evidence. According to the DIBELS-8 technical manual (University of Oregon, 2020), DIBELS-ORF has evidence of concurrent and delayed alternate form reliability and test-retest reliability. Delayed alternate form reliability has been reported to range from $r = .82$ to $.90$ for grade 4 and $r = .81$ to $.90$ for grade 5 (University of Oregon, 2020). Concurrent

alternate form reliability has been reported to range from $r = .81$ to $.95$ for grade 4 and $r = .89$ to $.95$ grade 5 (University of Oregon, 2020). Further, test-retest reliability has been reported to range from $r = .88$ to $.91$ for grade 4 and $r = .87$ to $.91$ for grade 5; the time between the two assessments was approximately three months (University of Oregon, 2020).

Concurrent validity in the form of correlations between DIBELS-ORF with a similar subtest from previous versions of DIBELS (University of Oregon, 2020) has been reported to be $r = .90$ for grade 4 and $r = .95$ for grade 5 (University of Oregon, 2020). Predictive correlations between DIBELS-ORF and the Iowa State Assessment (Iowa Testing Program, n.d.) were reported to range from $r = .53$ to $.69$ for grade 4 and $r = .46$ to $.63$ for grade 5 at the beginning of the academic year (University of Oregon, 2020). Criterion-referenced sensitivity values (correctly identifying students meeting standard) were reported to range from $.60$ to $.66$ for grade 4 and $.71$ to $.78$ for grade 5 (University of Oregon, 2020); specificity values (correctly identifying students not meeting standard) were reported to range from $.76$ to $.81$ for grade 4 and $.65$ to $.76$ for grade 5 (University of Oregon, 2020).

Data Analysis Plan

General Modeling Approach

The research questions for this study were evaluated with 2-level linear regression models estimated using full information maximum likelihood fitted by the *R* lme4 package (Bates et al., 2015). Satterthwaite degrees of freedom for regression coefficients were obtained using the lmerTest package (Kuznetsova et al., 2017). Multilevel modeling (also known as “random” and “mixed” modeling) affords the use of random, rather than fixed, effects for clustering variables (i.e., schools) so that lower-level effects can be separated between and within schools; as well, when upper-level effects are evaluated, they use the proper degrees of freedom

and standard errors. Following recommended practice (Luo et al., 2021), preliminary intercept-only models were specified before formal analyses to assess the degree of non-independence in scores due to clustering. Thereafter two sets of models were specified with students treated as L1 units nested within school L2 units (described next).

Across all models for ease of model results interpretation and prior to cluster-mean-centering, the three reading screeners were z-scored and the two binary predictors were effect-coded (grade level coded 1 = Gr. 5, -1 = Gr. 4; learner assistance program (LAP) status coded 1 = yes, -1 = no). A 2-tailed alpha of .05 was used for all analyses. Fit indices, along with coefficients and their effect sizes, are reported for each model result. Effect sizes included total fixed effects percent of variance-explained R^2 (Rights and Sterba, 2019) as well as the squared semi-partial correlations for each slope coefficient (unique percent of variance-explained). Effect size values were obtained using the *R* “MLMES” package (Sanders et al., 2026). Finally, model-based data visualization was implemented using *R* “ggplot2” (Wickham, 2016), and *R* “effects” packages (Fox & Weisberg, 2019).

Research Questions 1 and 2: Predictive Utility of Universal Screeners and Moderators

The first set of models (Model 1) included grade, LAP status, and *one* of each of the reading screener measures: (a) prior year SBA-ELA, (b) fall iReady Reading Diagnostic, or (c) fall DIBELS-ORF. Interactions among predictors were included as well for evaluating Research Question 2. Due to model complexity and the limited number of L2 units, only random intercepts (not random slopes) and student-level predictor effects were estimated. Model 1 was as follows.

$$\begin{aligned}
 SBA_ELA_{ij} = & \gamma_{00} + \gamma_{10}Grade_{ij} + \gamma_{20}LAP_{ij} + \gamma_{30}ZReadingScreener_{ij} + \\
 & \gamma_{40}Grade \times LAP_{ij} + \\
 & \gamma_{50}Grade \times ZReadingScreener_{ij} + \\
 & \gamma_{60}LAP \times ZReadingScreener_{ij} + \\
 & \gamma_{70}Grade \times LAP \times ZReadingScreener_{ij} + \\
 & U_j + R_{ij}
 \end{aligned}$$

In the model above, the i th spring 2024 SBA-ELA score for a student in the j th school a function of the sum of the conditional mean spring score (intercept, γ_{00}), the main effects of L1 predictors ($\gamma_{10} - \gamma_{30}$), the 2-way interaction terms among L1 predictors ($\gamma_{40} - \gamma_{60}$), the 3-way interaction term (γ_{70}), the school-level residual between the school's mean SBA-ELA and the conditional grand mean SBA-ELA score (i.e., random intercepts, U_j) and the within-school (student-level) residual (R_{ij}). Model-based follow-up simple slopes t -tests were used for evaluating differences in slopes for significant interactions.

Research Question 3: Unique Predictive Utility of Universal Screeners

For evaluating Research Question 3, Models 1a – 1c were combined three ways for comparison with Model 1a (i.e., the model with just the prior year's SBA-ELA as the reading screener predictor along with the grade and LAP status covariates and interactions). Model 2a combined prior year SBA-ELA with iReady Reading; Model 2b combined prior year SBA-ELA with DIBELS-ORF, and Model 2c combined all three reading screeners together. Likelihood ratio tests were used to test model fit differences, and pairwise follow-up t -tests were conducted using Model 2c's reading screener coefficient slope coefficients and corresponding coefficient variance-covariance elements to test for the significance of the differences among the three reading measures' unique effects.

Chapter 4: Results

Descriptive Statistics

Means, standard deviations, and zero-order correlations among variables are presented in Table 3. As can be seen, the dependent variable, spring 2024 SBA-ELA, averaged $M = 2531.81$ points ($SD = 96.95$) and the mean of the prior year's SBA-ELA was $M = 2484.22$ ($SD = 23.14$), indicating many students in the district met the state's reading standards (the state's SBA-ELA standard that year was 2473 for grade 4 and 2502 for grade 5) and that there was close to 50 points in growth for the sample from 2023 to 2024. The average fall 2023 iReady Reading was $M = 567.29$ ($SD = 10.91$), indicating that many students scored at least at the grade 4 placement level (Curriculum Associate, 2026c). The average score on DIBELS-ORF was $M = 115.13$ ($SD = 9.80$), indicating that many students demonstrated minimal risk of later reading difficulties (University of Oregon, 2023a).

The correlations in Table 3 show that all three measures were significantly and highly correlated with the dependent variable, spring 2024 SBA-ELA: at the student level (within schools), prior year SBA-ELA correlated at $r = .74$, fall iReady Reading correlated at $r = .77$, and DIBELS-ORF correlated at $r = .60$. At the school level (means), the three correlations were .20, .19, and .18, respectively. As can also be seen, the three measures were all highly correlated with each other: at the student level, the range was .65 – .81, and at the school-level, the range was .91 – .94 (i.e., nearly perfectly collinear; notably, school-level variables were not used as predictors for this reason as well as because there would be little power to test their effects given the small number of schools in the sample).

Grade level and LAP status were also predictive of the spring SBA-ELA scores: Grade 5 students were higher than Grade 4 students ($r = .22$) and LAP students were lower than their

peers ($r = -.32$), and this pattern of effects also holds for the three reading screener measures as well (prior year SBA-ELA, iReady Reading Diagnostic, and DIBELS-ORF).

Intraclass Correlations

Intercept-only (“empty”) models were specified to obtain the intraclass correlations (ICCs) for the dependent variable as well as the three focal reading predictors. An ICC is the degree of dependency among scores due to clustering and is computed as the between-school variance estimate divided by the total variance estimate. Those results showed that the ICC for the dependent variable, spring 2024 SBA-ELA, was .05, indicating that 5% of the variability in student scores was attributable to between-school differences (an alternative way of interpreting the ICC is also the expected correlation between any pair of randomly drawn scores within the same cluster, so in this case, the expected correlation between any pair of randomly drawn SBA-ELA scores within a given school was predicted to be $r = .05$). The reading measure predictors showed nearly identical results: ICCs for the prior year SBA-ELA, iReady Reading, and DIBELS-ORF were .06, .05, and .07, respectively; LAP status similarly had an approximate ICC of .06. In larger scale studies with numerous elementary schools, ICCs of .20 can be common (Hedges & Hedberg, 2007); in this study, all 10 schools were from the same district so less variability may be expected (as was found).

Research Question 1: Direct Predictive Utility of Reading Screeners

Table 4 displays the results for Model 1 (1a – 1c in the first, second, and third set of columns, respectively). These models were specifically used to test the *direct* predictive utility of each of the three reading measures on spring 2024 SBA-ELA scores, controlling for grade level and LAP status (and their interactions with each other). From the total R^2 results, we see that the sets of predictors explained a large share of variance in the SBA-ELA scores: the model with fall

iReady Reading was predicted to explain 58.7% of the variance in scores, and the model with prior year SBA-ELA as the reading screener was predicted to explain 54.3% of the variance. The model with DIBELS-ORF, on the other hand, only explained 38.2% of the variance.

As can be seen from the fixed effects coefficient results, fifth graders tended to score higher than fourth graders, but only significantly so in the model that employed DIBELS-ORF as the reading screener and students identified for LAP services tended to be lower than those not identified, but only significantly so for the model that employed prior year SBA-ELA as the reading screener. There was also one significant grade-by-LAP interaction for the model including DIBELS-ORF as the reading screener ($p = .014$), indicating that there was a significant negative LAP effect for fifth graders ($diff = -29.80$, $t(1198) = -2.59$, $p = .010$), but not fourth graders ($diff = 12.00$, $t(1193) = 0.94$, $p = .346$).

More interestingly, controlling for grade and LAP factors, all three reading screeners were significantly, positively associated with the spring 2024 SBA-ELA (p -values $< .001$). Prior year SBA-ELA was predicted to explain 40.2% of the variance in student scores, iReady Reading was predicted to explain 44.2% of the variance in student scores, and DIBELS-ORF was predicted to explain 21.5% of the variance in student scores. In other words, iReady Reading had the strongest relationship with spring SBA-ELA, but the magnitude was quite close to that of the prior year SBA-ELA's relationship. DIBELS-ORF, in contrast, was predicted to explain only 21.5% of the variance in spring SBA-ELA scores. Holding grade and LAP status constant, Figure 1 illustrates the model-implied relationships between each reading screener and the spring SBA-ELA: while prior year SBA-ELA and iReady Reading had nearly indistinguishable slopes, DIBELS-ORF's slope was markedly weaker than the other two screeners.

Research Question 2: Grade and LAP Effects on Screener Relations

Results shown in Table 4 also revealed that there were no significant interactions between the three screeners and grade or LAP, indicating that these factors did not moderate the direct predictive utility of the screeners. This said, a trend for a 3-way interaction was observed for DIBELS-ORF ($p = .066$). In Figure 2, the visual pattern of the interaction appears to indicate that the DIBELS-ORF predictive utility was weaker for fourth-grade students not identified for LAP compared to those that were identified for LAP, but follow-up tests indicated no significant in slope strength among the four subgroups (grade 4 LAP, grade 4 non-LAP, grade 5 LAP, grade 5 non-LAP): all six Model-based t -test p -values unadjusted for multiple comparisons $> .05$.

Research Question 3: Unique Predictive Utility of Reading Screeners

To evaluate the third research question, a second set of models (Models 2a – 2c) was specified that sequentially combined each of the first sets of models to determine the additive contributions of the iReady Reading and DIBELS-ORF screeners above and beyond students' prior year SBA-ELA for predicting spring 2024 SBA-ELA. Model 2a given in the first set of columns of Table 5 (combining prior year SBA-ELA and iReady Reading screeners together) fit the data significantly better than Model 1a (with just SBA-ELA as reading screener), $LRT \chi^2(4) = 263.61, p < .001$. The variance explained improved 10%, from 54.3% to 62.3%.

Model 2b given in the second set of columns of Table 5 (combining prior year SBA-ELA and DIBELS-ORF screeners together) also fit the data significantly better than Model 1a (with just SBA-ELA as reading screener), $LRT \chi^2(4) = 73.01, p < .001$. However, the total R^2 for this model was 56.7%, an improvement of just 2.4% additional variance explained beyond the model that just used the prior year's SBA-ELA.

Model 2c given in the third set of columns incorporated all three reading screeners together. As can be seen from the fixed effect coefficient results, although each screener did uniquely predict the outcome, the *unique* variance explained by each of the screeners was small because they shared so much variance with each other. The sr^2 values indicate that iReady Reading was the most uniquely predictive measure, explaining 5.4% of the variance in spring scores; prior year SBA-ELA uniquely explained 3.2% of the variance, and DIBELS-ORF uniquely explained only 0.2% of the variance. Using Model 2c's results, follow-up pairwise slope tests comparing the three measures' relationship strength showed that iReady Reading's slope was significantly greater than prior year SBA-ELA's slope, $t(1181) = 2.16, p = .018$, as well as DIBELS-ORF's slope, $t(1181) = 8.05, p < .001$; SBA-ELA's slope was also significantly greater than that for DIBELS-ORF, $t(1181) = 6.13, p < .001$.

This finding directly matches the final set of likelihood ratio test model comparisons: whereas the final model, Model 2c (with all three reading screeners), was a significant improvement over Model 2b (with SBA-ELA and DIBELS-ORF: LRT $\chi^2(4) = 198.16, p < .001$), it was not a significant improvement over Model 2a (with SBA-ELA and iReady Reading: LRT $\chi^2(4) = 7.57, p = .109$). In sum, prior year SBA-ELA and fall iReady Reading had similar predictive utility (but with iReady slightly better), whereas DIBELS-ORF did not contribute meaningfully to the prediction of end-of-year SBA-ELA more than the other reading screeners.

Chapter 5: Discussion

Universal screening is a widely used educational practice that serves as a core component of multi-tiered system of supports (MTSS) implementation. Although many universal screening measures are available, measures vary in terms of psychometric quality and feasibility. The purpose of this study was to compare the predictive utility – direct and unique – of three universal screening measures of grade 4-5 prior reading indicators on spring reading within one school district. These measures included the Smarter Balanced Assessment – English Language Arts (SBA-ELA) scores from the prior spring of 2023, the iReady Reading Diagnostic scores from fall of 2023, and the Dynamic Indicators of Basic Early Literacy Skills – Oral Reading Fluency (DIBELS-ORF) scores from fall of 2023. Additionally, the extent to which grade level and learner assistance program (LAP) status moderated these relationships was also examined.

While the iReady Reading Diagnostic's concurrent validity has been established with the SBA-ELA and other state assessments in earlier research (Curriculum Associates, 2020a, 2022a, 2022b, 2022c, 2022d), these studies were funded by the test developer and few examined the predictive utility of the iReady Reading Diagnostic assessment on future reading performance, both of which may be considered problematic since the iReady Reading Diagnostic is fairly expensive and quite commonly used (Benson et al., 2019). Further, no published studies have compared the predictive validity of iReady Reading Diagnostic with prior year SBA-ELA or DIBELS-ORF within the same sample.

As well, while the DIBELS-ORF is free and relatively easy to implement, its predictive utility has not been compared to other measures that schools and districts may already have and use. In other words, it is not known whether adding another layer of testing – albeit brief – is a good use of time and energy resources for teachers and students alike. Taken together, this study

contributes to the existing literature on universal reading screening by being able to directly compare three measures that one school district was already using to help guide their reading assessment choices.

Research Question 1: Direct Predictive Utility of Reading Screeners

Model results indicated that the iReady Reading Diagnostic scores from fall of 2023 were similarly good at predicting performance on spring of 2024 SBA-ELA as compared to the prior year's scores, but that iReady's relationship was stronger than that for fall 2023 DIBELS-ORF. This may be largely due to the content covered in these assessments. Specifically, the iReady Reading score is a composite of multiple subtests, as is the SBA-ELA. Depending on the grade level and performance of the student, the iReady subtests could include Vocabulary, Comprehension Literature, Comprehension Informational, Phonics, and High Frequency Words (NCII, n.d.). The inclusion of multiple reading domains is likely to have enhanced the predictive value of iReady Reading compared to DIBELS-ORF, which only assesses passage reading accuracy and fluency.

In fact, although DIBELS does offer composite scores (University of Oregon, 2023b), the district – like many – only use the DIBELS ORF in grades 4-5. One study found that oral reading fluency measures such as passage reading fluency (University of Houston, 2008), Test of Word Reading Efficiency (TOWRE; Torgesen et al., 1999) and word reading fluency (University of Houston, 2008) may be inadequate screening measures when used alone, resulting in only moderate levels of sensitivity and specificity (Denton et al., 2011). The inclusion of multiple subtests may have improved the predictive validity of iReady Reading above DIBELS-ORF. This finding is consistent with other studies that found DIBELS-ORF may have limited positive predictive accuracy (Roehrig et al., 2008).

Fewer studies have examined the predictive utility of iReady Reading. However, Campaña and Solomon (2025) found that iReady Reading showed evidence of sensitivity, with sensitivity values ranging from .68 to .90. These findings are consistent with research conducted on other CATs (Sutter et al., 2020; Thomas & January 2021). The present study adds to this body of literature by using multilevel modeling to account for clusters within the dataset, comparing iReady Reading and DIBELS-ORF directly, and using a larger sample size (1,191 students) to conclude that iReady Reading Diagnostic scores were a significantly better predictor of later state test performance than DIBELS-ORF scores.

Research Question 3: Unique Predictive Utility of Reading Screeners

Jumping forward to research question 3 (I will return to question 2 momentarily): after combining screeners into one model, results showed that DIBELS-ORF did not contribute significantly to the prediction of the end-of-year SBA-ELA scores beyond what the prior year and what iReady Reading could predict. In fact, iReady Reading was uniquely, significantly better at predicting spring SBA-ELA compared to prior year scores as well. While the DIBELS-ORF finding was not surprising given what was discussed above, the fact that iReady Reading was better at predicting spring SBA-ELA than the prior SBA-ELA may be due to the time gap between the two screeners—some students in grades 4-5 may have received summer enrichment that changed their scores between spring 2023 and fall 2023, thereby making the fall iReady Reading a better predictor given its 3-month closer proximity to the dependent variable. Again, as alluded to above, the SBA-ELA and iReady Reading share common content and both were nearly indistinguishable in their slopes in Figures 1-2.

In general, the findings for both Research Questions 1 and 3 are consistent with findings from Denton and colleagues (2011), Paly and colleagues (2022), and Campaña and Solomon

(2025). Denton and colleagues (2011) compared oral reading fluency curriculum-based measures (CBMs) to scores on the 2006 administration of the Texas Assessment of Knowledge and Skills (TAKS; Texas Education Association, 2004) on performance on the 2007 TAKS administration, finding that TAKS scores from 2006 yielded a classification accuracy value of .82, while the CBMs yielded classification accuracy values ranging from .69 to .72 (Denton et al., 2011).

In short, there may be little benefit to using both CBMs and past state assessment scores to predict student performance on future state assessments. Additionally, schools that use multiple measures for screening purposes, as was the case for the schools in this study, may find that past state test scores provide sufficient information in predicting future state test scores. Administering additional screening assessments to all students may not be a valuable use of time, as studies have found that combining screening approaches may not improve classification accuracy (Campaña & Solomon, 2025).

More recently, Paly and colleagues (2022) compared several approaches of universal screening of reading, including using prior state assessment scores, using a combination of state assessment scores and scores from aimswebPlus (NCS Pearson, 2018), a multiple gating model, and aimswebPlus scores alone. The study found that using prior state assessment scores as the first gate, with CBM scores as the second gate, yielded the highest overall classification accuracy (Paly et al., 2022). However, cost-effectiveness analyses indicated that the aimswebPlus only approached yielded the highest cost (i.e., \$55,198.52 total cost for elementary school students) and lowest effectiveness (i.e., highest number of false positive), while the prior state test results had high classification accuracy without additional financial cost (Paly et al., 2022). Several other studies demonstrated the predictive utility of prior state assessment scores in predicting later state assessment performance (Neslon et al., 2017; VanMeveren et al., 2020; Van Norman

et al., 2017), indicating that examining past state test data may be a viable and cost-effective option for universal screening. Previous state test scores have additional benefits as these scores would not require additional administration time or financial cost to use for universal screening.

Research Question 2: Grade and LAP Effects on Screener Relations

Model results examining interactions between grade level and LAP identification and the predictive utility of the three reading screeners suggested little in the way of meaningful moderation. In Model 1 with DIBELS-ORF, grade, and LAP status, follow-up tests of the 3-way interaction showed no significant subgroup differences in DIBELS-ORF's predictive utility. No other screener interactions were found despite the models being well-powered with a large student sample size. This null finding indicates good news and bad news: the good news is that the utility of the screeners worked similarly for grades 4 and 5 alike. The not-so-good news is that we might expect to see weaker predictive utility for students designated for LAP services, since those services are aimed at intervening and supporting students to increase their scores, but this was not observed. This finding could be due to several reasons. First, students' LAP designation is broad and includes an array of disabilities that may not be directly related to reading skills. Second, being designated as LAP does indicate the nature or strength of the type of service, if any, that was received. Third, it may be the case that students received support that helped them keep up with their previous skill level (i.e., not falling further behind) but not to the extent that sufficiently raised their scores beyond what their previous scores predicted for them.

Last, this finding may be due to the way data is used as many teachers report having little knowledge of assessment and data-based decision making (Al Otaiba et al., 2019; Spear-Swerling & Cheesman, 2012). Difficulties with using reading screening data to inform instruction could be influenced by availability of professional development, the amount of time

allotted for universal screening, and the use of formal decision-making procedures (Komesidou et al., 2022; Ozernov-Palchik et al., 2025; Stecker et al., 2005). In a survey of school staff members involved in universal screening for reading processes, Ozernov-Palchik and colleagues (2025) found that almost half of the participants reported a lack of formal procedures for developing intervention plans based on reading screening data. Additionally, half of the participants indicated that many students do not receive appropriate instruction after being identified during the universal screening process, and others reported feeling unprepared to design intervention plans (Ozernov-Palchik et al., 2025). Similarly, Hoffman and colleagues (2009) noted that some educators in their study appeared to be unsure of how to use DIBELS data to inform reading interventions. Other studies have highlighted the importance of adequate and ongoing professional development for using reading data effectively (Komesidou et al., 2022; Spear-Swerling & Cheesman, 2012; Stecker et al., 2005), with some educators stating that truncated professional development was a barrier to effective implementation of universal screening for reading (Komesidou et al., 2022).

While it is unclear how much professional development LAP interventionists in this study received, the results could indicate that additional professional development related to universal screening and data-based decision making in reading is warranted for the district. Another factor that could influence the usefulness of universal screening results is the fidelity of assessment administration. Reed and Sturgess (2013) found correctable administration errors in 91% of an ORF measure, which could impact instructional decision making. Unfortunately, data on administration fidelity were not available for the present study, representing a limitation. Fidelity data for DIBELS-ORF may be especially useful as it required teachers to administer the assessment to individual students.

Limitations and Future Directions

Results from this study should be interpreted in the context of limitations related to both internal, external, and statistical conclusion validity. Internal validity is the degree to which a study can demonstrate a causal relationship between predictor and outcome variables whereas external validity refers to the extent to which the study's results can be generalized to other settings, populations, and variables (Campbell, 1957; Shadish et al., 2001).

Threats to Internal Validity

First and foremost, although there is temporal evidence for prediction, this study is correlational in nature and no causal conclusions can be drawn. We cannot know if the screener content, screener administration, or scoring procedures (or some combination) caused the differences in predictive utility that we observed. Even with rigorous training, errors in assessment administration can occur (Reed & Sturges, 2013). Such errors can impact the inferences that can be drawn from universal screening data, making it difficult to use these data for educational decision making (Christ et al., 2012). Future work in this area should involve the collection of fidelity data, which could then be included as a predictor variable in subsequent models. For DIBELS-ORF, fidelity data related to adherence to testing scripts, timing procedures, and scoring procedures would provide information about the quality of each test administration. For iReady Reading, information on technological difficulties, on- and off-task student behavior, and teacher response to off-task student behaviors could provide information on assessment fidelity.

Threats to External Validity

The data from this study were drawn from a district that is mostly White and English-speaking (85%); as such, the homogeneity of this sample may impact the generalizability of

these results to other districts and states with more diverse student populations. For example, it should not be assumed that the results would hold true for districts with higher percentages of multilingual students. Relatedly, these universal screening data are from only one school district. The results of this study could have been impacted by policies, curriculum, testing procedures, and instructional resources specific to the local school district. This a threat that Shadish and colleagues (2001) referred to as the interaction of the causal relationship with settings. Future research could focus on replicating this study's finding with multiple school districts. It may be particularly beneficial to examine the predictive utility of DIBELS-ORF, iReady Reading, and past scores on the SBA-ELA within more culturally and linguistically diverse school districts.

Conclusion

Decades of national data indicate that many U.S. students struggle to read proficiently. Universal screening within an MTSS framework focuses on assessing the health of universal instruction, identifying students who would benefit from supplemental instruction, and allocating resources effectively to ensure that all students have access to high quality instruction. However, the efficacy of universal screening practices is dependent on the psychometric utility of the screening measures being used.

This study examined how well three measures of reading predict performance on an end-of-year assessment. These measures included prior year SBA-ELA scores, iReady Reading scores, and DIBELS-ORF. While all three measures were predictive of performance on the SBA-ELA from spring of 2024, DIBELS-ORF contributed little to the prediction of SBA-ELA scores. As well, although iReady Reading had the strongest predictive relationship among the three measures, the prior year's SBA-ELA scores were nearly as predictive (see again Table 4 and Figure 1). As a mandated state assessment, no additional administration time to obtain SBA-ELA

score from 2023 was required because that assessment had already been administered. Each of the assessments examined in this study had unique benefits. DIBELS-ORF only takes 1 to 3 minutes to administer and has evidence supporting its use for universal screening. iReady Reading provides information on important subskills that contribute to reading development. Additionally, iReady Reading can be group administered, items are presented using pre-recorded audio, and performance is scored automatically, requiring less logistical burdens for educators. Finally, access to past state assessment scores require no additional administration time, allowing educators to allocate more time to instruction.

The optimal reading screening measure will likely depend on the goals of the district, available resources for administration, and the type of assessment questions being answered. For example, if a district is interested in knowing which students are on track to pass the next state assessment, then prior assessment score data may be a useful measure. However, if an educator wants to design a reading intervention program for students, then iReady Reading may be useful because it measures important reading subskills such as phonological awareness, phonics, vocabulary, and comprehension. If monitoring progress over time is the priority, then DIBELS-ORF may be useful because it is quick to administer, contains multiple forms, and is sensitive to growth. In other words, the purpose of the assessment should inform the type of measure that is used.

References

- Allington, R. L. (1983). Fluency: The neglected reading goal. *The Reading Teacher*, 36(6), 556–561.
- Al Otaiba, S., Baker, K., Lan, P., Allor, J., Rivas, B., Yovanoff, P., & Kamata, A. (2019). Elementary teacher's knowledge of response to intervention implementation: a preliminary factor analysis. *Annals of Dyslexia*, 69(1), 34–53.
<https://doi.org/10.1007/s11881-018-00171-5>
- Altani, A., Georgiou, G. K., Deng, C., Cho, J.-R., Katopodi, K., Wei, W., & Protopapas, A. (2017). Is processing of symbols and words influenced by writing system? Evidence from Chinese, Korean, English, and Greek. *Journal of Experimental Child Psychology*, 164, 117–135. <https://doi.org/10.1016/j.jecp.2017.07.006>
- Altani, A., Protopapas, A., Katopodi, K., & Georgiou, G. K. (2020). From individual word recognition to word list and text reading fluency. *Journal of Educational Psychology*, 112(1), 22–39. <https://doi.org/10.1037/edu0000359>
- Anderson, C. M., & Borgmeier, C. (2010). Tier II interventions within the framework of school-wide positive behavior support: Essential features for design, implementation, and maintenance. *Behavior Analysis in Practice*, 3(1), 33–45.
<https://doi.org/10.1007/BF03391756>
- Ahmed, Y., Francis, D. J., York, M., Fletcher, J. M., Barnes, M., & Kulesz, P. (2016). Validation of the direct and inferential mediation (DIME) model of reading comprehension in grades 7 through 12. *Contemporary Educational Psychology*, 44–45, 68–82.
<https://doi.org/10.1016/j.cedpsych.2016.02.002>

- Ahmed, Y., Miciak, J., Taylor, W. P., & Francis, D. J. (2022). Structure altering effects of a multicomponent reading intervention: An application of the direct and inferential mediation (DIME) model of reading comprehension in upper elementary grades. *Journal of Learning Disabilities*, 55(1), 58–78. <https://doi.org/10.1177/0022219421995904>
- Ardoin, S. P., & Christ, T. J. (2008). Evaluating curriculum-based measurement slope estimates using data from triannual universal screenings. *School Psychology Review*, 37(1), 109–125. <https://doi.org/10.1080/02796015.2008.12087912>
- Aud, S. L., Fox, M. A., & KewalRamani, A. (2010). Status and trends in the education of racial and ethnic groups. *U.S. Dept. of Education, National Center for Education Statistics, Institute of Education Sciences*. <https://purl.fdlp.gov/GPO/gpo3616>
- Badian, N. A. (1999). Reading disability defined as a discrepancy between listening and reading comprehension: A longitudinal study of stability, gender differences, and prevalence. *Journal of Learning Disabilities*, 32(2), 138–148. <https://doi.org/10.1177/002221949903200204>
- Barrett, C. A., Johnson, L. J., Truckenmiller, A. J., & VanDerHeyden, A. M. (2024). Comparing the cost–accuracy ratios of multiple approaches to reading screening in elementary schools. *Remedial and Special Education*, 45(2), 71–84. <https://doi.org/10.1177/07419325231190809>
- Barron, S. I., & Koretz, D. M. (1995). An evaluation of the robustness of the national assessment of educational progress trend estimates for racial-ethnic subgroups. *Educational Assessment*, 3(3), 209–248. https://doi.org/10.1207/s15326977ea0303_1
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.

<https://doi.org/10.18637/jss.v067.i01>

Beck, I. L., McKeown, M. G., & Kucan, L. (2013). *Bringing Words to Life: Robust Vocabulary Instruction. Second Edition*. Guilford Press.

Beck, I. L., Perfetti, C. A., & McKeown, M. G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology*, 74(4), 506–521. <https://doi.org/10.1037/0022-0663.74.4.506>

Bell, N., & Wheldall, K. (2022). Factors contributing to reading comprehension in children with varying degrees of word-level proficiency. *The Australian Journal of Education*, 66(1), 73–91. <https://doi.org/10.1177/00049441211062941>

Betts, J., Pickart, M., & Heistad, D. (2009). An investigation of the psychometric evidence of CBM-R passage equivalence: Utility of readability statistics and equating for alternate forms. *Journal of School Psychology*, 47(1), 1–17.
<https://doi.org/10.1016/j.jsp.2008.09.001>

Bhattacharya, A., & Ehri, L. C. (2004). Graphosyllabic analysis helps adolescent struggling readers read and spell words. *Journal of Learning Disabilities*, 37(4), 331–348.
<https://doi.org/10.1177/00222194040370040501>

Biemiller, A., & Boote, C. (2006). An effective method for building meaning vocabulary in primary grades. *Journal of Educational Psychology*, 98(1), 44–62.
<https://doi.org/10.1037/0022-0663.98.1.44>

Björn, P. M., Aunola, K., & Nurmi, J.-E. (2016). Primary school text comprehension predicts mathematical word problem-solving skills in secondary school. *Educational Psychology*, 36(2), 362–377. <https://doi.org/10.1080/01443410.2014.992392>

Blachowicz, C. L. Z., Fisher, P. J. L., Ogle, D., & Watts-Taffe, S. (2006). Vocabulary: Questions

from the classroom. *Reading Research Quarterly*, 41(4), 524–539.

<https://doi.org/10.1598/RRQ.41.4.5>

Braun, H., Zhang, J., & Vezzu, S. (2010). An investigation of bias in reports of the national assessment of educational progress. *Educational Evaluation and Policy Analysis*, 32(1), 24–43. <https://doi.org/10.3102/0162373709351137>

Bruhn, A. L., Wehby, J. H., & Hasselbring, T. S. (2020). Data-based decision making for social behavior: Setting a research agenda. *Journal of Positive Behavior Interventions*, 22(2), 116–126. <https://doi.org/10.1177/1098300719876098>

Buckendahl, C. W., Davis, S. L., Plake, B. S., Sireci, S. G., Hambleton, R. K., Zenisky, A. L., & Wells, C. S. (2009a). *Evaluation of the National Assessment of Educational Progress*, Study Reports. U.S. Department of Education, Office of Planning, Evaluation and Policy Development, Washington, D.C., 2009.

<https://www.ed.gov/sites/ed/files/rschstat/eval/other/naep/naep-complete.pdf>

Buckendahl, C. W., Plake, B. S., & Davis, S. L. (2009b). Conducting a lifecycle audit of the national assessment of educational progress. *Applied Measurement in Education*, 22(4), 321–338. <https://doi.org/10.1080/08957340903221642>

Burns, M. K., VanDerHeyden, A. M., Duesenberg-Marshall, M. D., Romero, M. E., Stevens, M. A., Izumi, J. T., & McCollom, E. M. (2023). Decision accuracy of commonly used dyslexia screeners among students who are potentially at risk for reading difficulties. *Learning Disability Quarterly*, 46(4), 306–316.

<https://doi.org/10.1177/07319487221096684>

Cabell, S. Q., Gerde, H. K., Hwang, H., Bowles, R., Skibbe, L., Piasta, S. B., & Justice, L. M. (2022). Rate of growth of preschool-age children’s oral language and decoding skills

- predicts beginning writing ability. *Early Education and Development*, 33(7), 1198–1221.
<https://doi.org/10.1080/10409289.2021.1952390>
- Campaña, K. V., & Solomon, B. G. (2025). Classification accuracy of i-Ready and prior year state exams on year-end outcomes. *Assessment for Effective Intervention*, 50(3), 153–162.
<https://doi.org/10.1177/15345084251338860>
- Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54(4), 297–312. <https://doi.org/10.1037/h0040950>
- Casey, A., Deno, S., Marston, D., & Skiba, R. (1988). Experimental teaching: Changing beliefs about effective instructional practices. *Teacher Education and Special Education*, 11(3), 123–131. <https://doi.org/10.1177/088840648801100305>
- Catts, H. W., Petscher, Y., Schatschneider, C., Sittner Bridges, M., & 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>
- Cho, E., & Barrett, C. A. (2024). Comparing options for screening of reading difficulties in middle school: Do teacher ratings improve accuracy? *School Psychology*, 40(5), 542–553. <https://doi.org/10.1037/spq0000674>
- Christ, T. J., Zopluoglu, C., Long, J., & Monaghan, B. (2012). Curriculum-based measurement of oral reading: Quality of progress monitoring outcomes. *Exceptional Children*, 78(3), 356–373. <https://doi.org/10.1177/001440291207800305>
- Clemens, N. H., Hagan-Burke, S., Luo, W., Cerda, C., Blakely, A., Frosch, J., Gamez-Patience, B., & Jones, M. (2015). The predictive validity of a computer-adaptive assessment of

- kindergarten and first-grade reading skills. *School Psychology Review*, 44(1), 76–97.
<https://doi.org/10.17105/SPR44-1.76-97>
- Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review*, 24(4), 377–392.
<https://doi.org/10.1016/j.econedurev.2004.06.008>
- Colorado Department of Education. (2021). *CMAS – Mathematics, English language arts, science, and social studies assessments*. <https://www.cde.state.co.us/assessment/cmas>
- Compton, D. L., Fuchs, D., Fuchs, L. S., Bryant, J. D., & Harris, K. R. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*, 98(2), 394–409.
<https://doi.org/10.1037/0022-0663.98.2.394>
- Cooc, N. (2023). National trends in special education and academic outcomes for English learners with disabilities. *The Journal of Special Education*, 57(2), 106–117.
<https://doi.org/10.1177/00224669221147272>
- Coyne, M. D., McCoach, D. B., Loftus, S., Zipoli, R., & Kapp, S. (2009). Direct vocabulary instruction in kindergarten: Teaching for breadth versus depth. *The Elementary School Journal*, 110(1), 1–18. <https://doi.org/10.1086/598840>
- Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, 99(2), 311–325.
<https://doi.org/10.1037/0022-0663.99.2.311>
- Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010). Reading comprehension of scientific text: A domain-specific test of the direct and inferential mediation model of

reading comprehension. *Journal of Educational Psychology*, 102(3), 687–700.

<https://doi.org/10.1037/a0019452>

Cummings, K. D., Park, Y., & Bauer Schaper, H. A. (2013). Form effects on DIBELS next oral reading fluency progress- monitoring passages. *Assessment for Effective Intervention*, 38(2), 91–104. <https://doi.org/10.1177/1534508412447010>

Curriculum Associates. (n.d.). *iReady Diagnostic Reading*.

<https://www.curriculumassociates.com/reviews/assessment/diagnostic-scores-placements>

Curriculum Associates. (n.d.b.). *Frequently Asked Questions about iReady*.

<https://www.curriculumassociates.com/family>

Curriculum Associates. (n.d.c.). *iReady Assessment for the Purpose of Instruction*.

https://ebooks.curriculumassociates.com/story/assessment-for-instruction/page/1?pid=61954e2519983555f5e057f3&_gl=1*_lqunuun*_gcl_au*NTAzOTg2NTIuMTc0MjUxNTY1OA..*_ga*MjA0NDMxMDk2NS4xNzQyNTE1NjU5*_ga_64DEB7D9ZT*MTc0MjUyMDIwOS4yLjEuMTc0MjUyMDU1OC4wLjAuMA..*_ga_RTBB90JLP5*MTc0MjUyMDIwOS4yLjEuMTc0MjUyMDU1OC4wLjAuNTU2NTUxNTU2

Curriculum Associates (2020a). *The Relationship between i-Ready® Diagnostic and the 2019 Mississippi Academic Assessment Program (MAAP)*.

<https://cdn.bfldr.com/LS6J0F7/at/wtw9g6kbpwmwhnfrfx37zv/iready-full-correlation-brief-mississippi.pdf>

Curriculum Associates (2020b). *Using Lexile Measures from the Diagnostic Results Report*.

<https://cdn.bfldr.com/LS6J0F7/at/39pqm5wz9cw8sn2v6jw583x/iready-metametrics-lexile-2020.pdf>

Curriculum Associates (2022a). *The Relationship between i-Ready Diagnostic and the 2021 Smarter Balanced Assessment (SBA)*.

<https://cdn.bfldr.com/LS6J0F7/at/t3fs5jmtx42x5hhx89ngphf/iready-diagnostic-full-correlation-brief-smarter-balanced-hawaii.pdf>

Curriculum Associates (2022b). *The Relationship between i-Ready Diagnostic and the 2022 Alabama Comprehensive Assessment Program (ACAP)*.

<https://cdn.bfldr.com/LS6J0F7/at/6f9prvgg8tr9hx77q5nj4t4w/iready-diagnostic-correlation-study-alabama-2022.pdf>

Curriculum Associates (2022c). *The Relationship between i-Ready Diagnostic and the Massachusetts Comprehensive Assessment System (MCAS)*.

<https://cdn.bfldr.com/LS6J0F7/at/w93rkt8s4x8p58nccwm3swq/iready-diagnostic-full-correlation-brief-massachusetts.pdf>

Curriculum Associates (2022d). *The Relationship between iReady Diagnostic and the 2022 Smarter Balanced Assessment (SBA)*.

<https://cdn.bfldr.com/LS6J0F7/at/phph6bqz8k8s54v9rwjkc96/iready-diagnostic-full-correlation-brief-smarter-balanced-assessment.pdf>

Curriculum Associates (2023a). *The Relationship between iReady Diagnostic and the 2022 Colorado Measures of Academic Success (CMAS)*.

<https://cdn.bfldr.com/LS6J0F7/as/h8xtg8hw9bgmb8nx7hs955r/iready-full-correlation-brief-colorado>

Curriculum Associates (2023b) *The Relationship between iReady Diagnostic and the 2022 Smarter Balanced Assessment (SBA) in Connecticut*.

<https://cdn.bfldr.com/LS6J0F7/as/gnmwqx8jkwg4wq239jwr/iready-diagnostic-full-correlation-brief-connecticut>

Curriculum Associates (2024). *New Hampshire Statewide Assessment System (NH SAS) Linking Study*. <https://www.curriculumassociates.com/research-and-efficacy/nh-sas-linking-study>

Curriculum Associates (2026a). *iReady Diagnostic National Norms Tables for Reading and Mathematics for Grades K–8*.

<https://cdn.bfldr.com/LS6J0F7/at/jx38srs8nbmnmvqxgwjxrj5/iready-diagnostic-norms-tables-K-8.pdf>

Curriculum Associates (2026b). *i-Ready Inform Assessment in Reading and Mathematics: What It Measures*. <https://cdn.bfldr.com/LS6J0F7/as/9kkp35gxsm3mffhv96rrp6n/iready-deep-dive-iready-diagnostic-what-it-measures>

Curriculum Associates (2026c). *i-Ready Inform Grades K-12 Scale Score Placement Tables*. <https://cdn.bfldr.com/LS6J0F7/as/s677f7rxj6cm37bt8npbgffp/iready-diagnostic-placement-tables>

Daane, M.C., Campbell, J.R., Grigg, W.S., Goodman, M.J., & Oranje, A. (2005). *Fourth-grade Students Reading Aloud: NAEP 2002 Special Study of Oral Reading* (NCES 2006-469). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. <https://nces.ed.gov/nationsreportcard/pubs/studies/2006469.aspx>

David, K., & Marchant, G. (2015). The achievement gaps in the United States: Race, poverty, and interactions over ten years. *The International Journal of Assessment and Evaluation*, 22(4), 1–15. <https://doi.org/10.18848/2327-7920/CGP/v22i04/48378>

- Decker, D. M., Hixson, M. D., Shaw, A., & Johnson, G. (2014). Classification accuracy of oral reading fluency and maze in predicting performance on large-scale reading assessments. *Psychology in the Schools*, 51(6), 625–635. <https://doi.org/10.1002/pits.21773>
- de Jong, P. F. (2011). What discrete and serial rapid automatized naming can reveal about reading. *Scientific Studies of Reading*, 15(4), 314–337. <https://doi.org/10.1080/10888438.2010.485624>
- Deke, J., Dragoset, L., Bogen, K., & Gill, B. (2012). *Impacts of Title I Supplemental Educational Services on Student Achievement* (NCEE 2012-4053). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. <https://ies.ed.gov/use-work/resource-library/report/evaluation-report/impacts-title-i-supplemental-educational-services-student-achievement>
- Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, 37(3), 184–192. <https://doi.org/10.1177/00224669030370030801>
- Deno, S. L., Fuchs, L. S., Marston, D., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review*, 30(4), 507–524. <https://doi.org/10.1080/02796015.2001.12086131>
- Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, 49(1), 36–47. <https://doi.org/10.1177/001440298204900105>
- Deno, S. L., & Mirkin, P. K. (1977). *Data-Based Program Modification: A Manual*. Exceptional Children's Program, Bureau of Adult and Occupational Education, U. S. Office of Education, Department of Health, Education, and Welfare. <https://files.eric.ed.gov/fulltext/ED144270.pdf>

- Deno, S. L., Reschly, A. L., Lembke, E. S., Magnusson, D., Callender, S. A., Windram, H., & Stachel, N. (2009). Developing a school-wide progress-monitoring system. *Psychology in the Schools*, 46(1), 44–55. <https://doi.org/10.1002/pits.20353>
- Denton, C. A., Barth, A. E., Fletcher, J. M., Wexler, J., Vaughn, S., Cirino, P. T., Romain, M., & Francis, D. J. (2011). The relations among oral and silent reading fluency and comprehension in middle school: Implications for identification and instruction of students with reading difficulties. *Scientific Studies of Reading*, 15(2), 109–135. <https://doi.org/10.1080/10888431003623546>
- Department of Education State of California. (n.d.). *Test Results at a Glance*. <https://caaspp-elpac.ets.org/caaspp/DashViewReportSB?ps=true&lstTestYear=2024&lstTestType=B&lstGroup=1&lstSubGroup=1&lstSchoolType=A&lstGrade=13&lstCounty=00&lstDistrict=00000&lstSchool=0000000>
- Dowhower, S. L. (1991). Speaking of prosody: Fluency's unattended bedfellow. *Theory Into Practice*, 30(3), 165–175. <https://doi.org/10.1080/00405849109543497>
- Duke, N. K., & Cartwright, K. B. (2021). The science of reading progresses: Communicating advances beyond the simple view of reading. *Reading Research Quarterly*, 56(S1), S25–S44. <https://doi.org/10.1002/rrq.411>
- Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading*, 9(2), 167–188. https://doi.org/10.1207/s1532799xssr0902_4
- Ehri, L., & Saltmarsh, J. (1995). Beginning readers outperform older disabled readers in learning to read words by sight. *Reading & Writing*, 7(3), 295–326. <https://doi.org/10.1007/BF03162082>

- Ehri, L. C., & Wilce, L. S. (1987). Cipher versus cue reading: An experiment in decoding acquisition. *Journal of Educational Psychology*, 79(1), 3–13.
<https://doi.org/10.1037/0022-0663.79.1.3>
- Elleman, A. M., & Oslund, E. L. (2019). Reading comprehension research: Implications for practice and policy. *Policy Insights from the Behavioral and Brain Sciences*, 6(1), 3–11.
<https://doi.org/10.1177/2372732218816339>
- Espin, C., Wallace, T., Lembke, E., Campbell, H., & Long, J. D. (2010). Creating a progress-monitoring system in reading for middle-school students: Tracking progress toward meeting high-stakes standards. *Learning Disabilities Research and Practice*, 25(2), 60–75. <https://doi.org/10.1111/j.1540-5826.2010.00304.x>
- Fahle, E. M., Shear, B. R., & Shores, K. A. (2019). Assessment for monitoring of education systems: The U.S. example. *The Annals of the American Academy of Political and Social Science*, 683(1), 58–74. <https://doi.org/10.1177/0002716219841014>
- Filderman, M. J., Toste, J. R., Didion, L. A., Peng, P., & Clemens, N. H. (2018). Data-based decision making in reading interventions: A synthesis and meta-analysis of the effects for struggling readers. *The Journal of Special Education*, 52(3), 174–187.
<https://doi.org/10.1177/0022466918790001>
- Flores, M., Poulos, J., Pennell, A., McKechnie, M., LaPalm, M., Toney, A. (2020). *Washington's Multi-Tiered System of Supports Framework*. Office of the Superintendent of Public Instruction. https://ospi.k12.wa.us/sites/default/files/2023-10/washington_s-mtss-framework-document.pdf
- Florida Department of Education. (n.d.). *2023-24 Florida Report Card*.
<https://edudata.fldoe.org/ReportCards/Schools.html?school=0000&district=00>

- Ford, J. W., Missall, K. N., Hosp, J. L., Kuhle, J. L. (2017). Examining oral passage reading rate across three curriculum-based measurement tools for predicting grade-level proficiency. *School Psychology Review*, 46(4), 363–378. <https://doi.org/10.17105/SPR-2016-0014.V46-4>
- Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression, 3rd Edition*. Sage Publications.
- Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review*, 21(1), 45–58. <https://doi.org/10.1080/02796015.1992.12085594>
- Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1994). Strengthening the connection between assessment and instructional planning with expert systems: Advances in technology-based assessment within special education. *Exceptional Children*, 61(2), 138–146.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review*, 22(1), 27–48. <https://doi.org/10.1080/02796015.1993.12085636>
- Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, 5(3), 239–256. https://doi.org/10.1207/S1532799XSSR0503_3
- Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education*, 9(2), 20–28. <https://doi.org/10.1177/074193258800900206>

- Georgiou, G. K., Parrila, R., Cui, Y., & Papadopoulos, T. C. (2013). Why is rapid automatized naming related to reading? *Journal of Experimental Child Psychology*, 115(1), 218–225.
<https://doi.org/10.1016/j.jecp.2012.10.015>
- Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology*, 45(2), 117–135.
<https://doi.org/10.1016/j.jsp.2006.05.005>
- Goffreda, C. T., & DiPerna, J.C. (2010). An empirical review of psychometric evidence for the dynamic indicators of basic early literacy skills. *School Psychology Review*, 39(3), 463–483. <https://doi.org/10.1080/02796015.2010.12087766>
- Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for 3rd grade high-stakes outcomes. *Scientific Studies of Reading*, 5(3), 257–288.
https://doi.org/10.1207/S1532799XSSR0503_4
- Goodrich, J. M., Thayer, L., & Leiva, S. (2021). Evaluating achievement gaps between monolingual and multilingual students. *Educational Researcher*, 50(7), 429-441.
<https://doi.org/10.3102/0013189X21999043>
- Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7(1), 6–10. <https://doi.org/10.1177/074193258600700104>
- Graves, M. F., Ringstaff, C., Li, L., & Flynn, K. (2018). Effects of teaching upper elementary grade students to use word learning strategies. *Reading Psychology*, 39(6), 602–622.
<https://doi.org/10.1080/02702711.2018.1496503>

Hanselman, P. (2019). Access to effective teachers and economic and racial disparities in opportunities to learn. *Sociological Quarterly*, 60(3), 498–534.

<https://doi.org/10.1080/00380253.2019.1625732>

Hattan, C., Lee, E., & List, A. (2023). Comprehension, diagram analysis, integration, and interest: A cross-sectional analysis. *Reading Psychology*, 44(7), 731–760.

<https://doi.org/10.1080/02702711.2023.2187907>

Haley, S. M., Chafetz, R. S., Tian, F., Montpetit, K., Watson, K., Gorton, G., & Mulcahey, M. J. (2010). Validity and reliability of physical functioning computer-adaptive tests for children with cerebral palsy. *Journal of Pediatric Orthopaedics*, 30(1), 71–75.

<https://doi.org/10.1097/BPO.0b013e3181c85453>

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1), 60–87. <https://doi.org/10.3102/0162373707299706>

Hemphill, F. C., & Vanneman, A. (2011). *Achievement Gaps: How Hispanic and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress* (NCES 2011-459). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

<https://nces.ed.gov/nationsreportcard/pdf/studies/2011459.pdf>

Hintze, J. M., & Silbergitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review*, 34(3), 372–386. <https://doi.org/10.1080/02796015.2005.12086292>

- Hoffman, A. R., Jenkins, J. E., & Dunlap, S. K. (2009). Using DIBELS: A survey of purposes and practices. *Reading Psychology, 30*(1), 1–16.
<https://doi.org/10.1080/02702710802274820>
- Hoogland, I., Schildkamp, K., van der Kleij, F., Heitink, M., Kippers, W., Veldkamp, B., & Dijkstra, A. M. (2016). Prerequisites for data-based decision making in the classroom: Research evidence and practical illustrations. *Teaching and Teacher Education, 60*, 377–386. <https://doi.org/10.1016/j.tate.2016.07.012>
- Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading & Writing, 2*(2), 127–160. <https://doi.org/10.1007/BF00401799>
- Hoover, W. A., & Tunmer, W. E. (2018). The simple view of reading: Three assessments of its adequacy. *Remedial and Special Education, 39*(5), 304–312.
<https://doi.org/10.1177/0741932518773154>
- Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABCs of CBM: A Practical Guide to Curriculum-Based Measurement (Second Ed.)*. Guilford Publications.
- Hudson, R. F., Pullen, P. C., Lane, H. B., & Torgesen, J. K. (2009). The complex nature of reading fluency: A multidimensional view. *Reading & Writing Quarterly, 25*(1), 4–32.
<https://doi.org/10.1080/10573560802491208>
- Hunley, S. A., Davies, S. C., & Miller, C. R. (2013). The relationship between curriculum-based measures in oral reading fluency and high-stakes tests for seventh grade students. *RMLE Online: Research in Middle Level Education, 36*(5), 1–8.
<https://doi.org/10.1080/19404476.2013.11462098>

Iowa Testing Programs. (n.d.). *Iowa Assessments, Forms E and F*.

<https://education.uiowa.edu/research/research-centers-and-research-initiatives/iowa-testing-programs>

January, S.-A. A., & Klingbeil, 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., & Fuchs, L. S. (2012). Curriculum-based measurement: The paradigm, history, and legacy. In C. A. Espin, K. L. McMaster, S. Rose, & M.M. Wayman (Eds.), *A Measure of Success: The influence of Curriculum-Based Measurement on Education* (pp. 7–23). University of Minnesota Press.

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

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>

Johnson, E. G. (1992). The design of the national assessment of educational progress. *Journal of Educational Measurement, 29*(2), 95–110. <https://doi.org/10.1111/j.1745-3984.1992.tb00369.x>

Jones, L. V. (1996). A history of the national assessment of educational progress and some questions about its future. *Educational Researcher, 25*(7), 15-22.

Jung, P.-G., McMaster, K. L., Kunkel, A. K., Shin, J., & Stecker, P. M. (2018). Effects of data-based individualization for students with intensive learning needs: A meta-

analysis. *Learning Disabilities Research and Practice*, 33(3), 144–155.

<https://doi.org/10.1111/ldrp.12172>

Kangas, S. E. N., Dai, S., & Ardasheva, Y. (2024). The intersection of language and disability progress of English learners with disabilities on NAEP reading. *The Journal of Special Education*, 58(2), 88–99. <https://doi.org/10.1177/00224669231213054>

Kaminski, R. A., & Good, R. H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review*, 25(2), 215–227.

<https://doi.org/10.1080/02796015.1996.12085812>

Kearns, D. M., & Hiebert, E. H. (2022). The word complexity of primary-level texts: Differences between first and third grade in widely used curricula. *Reading Research Quarterly*, 57(1), 255–285. <https://doi.org/10.1002/rrq.429>

Kelso, K., Whitworth, A., Parsons, R., & Leitão, S. (2022). Hidden reading difficulties: Identifying children who are poor comprehenders. *Learning Disability Quarterly*, 45(3), 225–236. <https://doi.org/10.1177/0731948720961766>

Kent, S. C., Wanzek, J., & Yun, J. (2019). Screening in the upper elementary grades: Identifying fourth-grade students at-risk for failing the state reading assessment. *Assessment for Effective Intervention*, 44(3), 160–172. <https://doi.org/10.1177/1534508418758371>

Kim, Y.-S. G. (2015). Developmental, component-based model of reading fluency: An investigation of predictors of word-reading fluency, text-reading fluency, and reading comprehension. *Reading Research Quarterly*, 50(4), 459–481.

<https://doi.org/10.1002/rrq.107>

Kim, Y.-S. G., & Wagner, R. K. (2015). Text (oral) reading fluency as a construct in reading development: An investigation of its mediating role for children from grades 1 to

4. *Scientific Studies of Reading*, 19(3), 224–242.

<https://doi.org/10.1080/10888438.2015.1007375>

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>

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2), 163–182. <https://doi.org/10.1037/0033-295X.95.2.163>

Klingbeil, D. A., McComas, J. J., Burns, M. K., & Helman, L. (2015). Comparison of predictive validity and diagnostic accuracy of screening measures of reading skills. *Psychology in the Schools*, 52(5), 500–514. <https://doi.org/10.1002/pits.21839>

Komesidou, R., Feller, M. J., Wolter, J. A., Ricketts, J., Rasner, M. G., Putman, C. A., & Hogan, T. P. (2022). Educators' perceptions of barriers and facilitators to the implementation of screeners for developmental language disorder and dyslexia. *Journal of Research in Reading*, 45(3), 277–298. <https://doi.org/10.1111/1467-9817.12381>

Kuder, S. J. (2017). Vocabulary instruction for secondary students with reading disabilities: An updated research review. *Learning Disability Quarterly*, 40(3), 155–164. <https://doi.org/10.1177/0731948717690113>

Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, 95(1), 3–21. <https://doi.org/10.1037/0022-0663.95.1.3>

- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82, 1-26.
<https://doi.org/10.18637/jss.v082.i13>
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6(2), 293–323. [https://doi.org/10.1016/0010-0285\(74\)90015-2](https://doi.org/10.1016/0010-0285(74)90015-2)
- Lane, S., Zumbo, B. D., Abedi, J., Benson, J., Dossey, J., Elliott, S. N., Kane, M., Linn, R., Paredes-Ziker, C., Rodriguez, M., Schraw, G., Slattery, J., Thomas, V., & Willhoft, J. (2009). Prologue: An introduction to the evaluation of NAEP. *Applied Measurement in Education*, 22(4), 309–316. <https://doi.org/10.1080/08957340903221436>
- Leach, J., Scarborough, H. S., & Rescorla, L. (2003). Late-emerging reading disabilities. *Journal of Educational Psychology*, 95(2), 211–224. <https://doi.org/10.1037/0022-0663.95.2.211>
- Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher*, 31(1), 3–12.
<https://doi.org/10.3102/0013189X031001003>
- Lee, J., & Reeves, T. (2012). Revisiting the impact of NCLB high-stakes school accountability, capacity, and resources: State NAEP 1990—2009 reading and math achievement gaps and trends. *Educational Evaluation and Policy Analysis*, 34(2), 209–231.
<https://doi.org/10.3102/0162373711431604>
- Lee, S. H. (2024). Effects of moderately rich vocabulary instruction on literacy performances of students with reading disabilities. *Communication Disorders Quarterly*, 45(4), 253–262.
<https://doi.org/10.1177/15257401231185526>

- Li, L., Zhu, D., & Wu, X. (2021). The effects of vocabulary breadth and depth on reading comprehension in middle childhood: The mediator role of listening comprehension. *Reading & Writing Quarterly*, 37(4), 336–347.
<https://doi.org/10.1080/10573569.2020.1809585>
- Liberman, I. Y., Shankweiler, D., Fischer, F. W., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology*, 18(2), 201–212. [https://doi.org/10.1016/0022-0965\(74\)90101-5](https://doi.org/10.1016/0022-0965(74)90101-5)
- Luo, T., Lee, G.-L., & Molina, C. (2017). Incorporating IStation into early childhood classrooms to improve reading comprehension. *Journal of Information Technology Education*, 16, 247–266. <https://doi.org/10.28945/3788>
- Luo, W., Li, H., Baek, E., Chen, S., Lam, K. H., & Semma, B. (2021). Reporting practice in multilevel modeling: A revisit after 10 years. *Review of Educational Research*, 91(3), 311–355. <https://doi.org/10.3102/0034654321991229>
- Maki, K. E., & Hammerschmidt-Snidarich, S. (2022). Reading fluency intervention dosage: A novel meta-analysis and research synthesis. *Journal of School Psychology*, 92, 148–165.
<https://doi.org/10.1016/j.jsp.2022.03.008>
- Majeika, C. E., Pierce, J., Smith, H., Lembke, E., & Gandhi, A. (2024). Integrated multi-tiered systems of support in elementary schools: Practical applications. *Intervention in School and Clinic*, 60(1), 53–61. <https://doi.org/10.1177/10534512241254031>
- Mason, J. M. (1980). When do children begin to read: An exploration of four year old children's letter and word reading competencies. *Reading Research Quarterly*, 15(2), 203–227.
<https://doi.org/10.2307/747325>

- McGlinchey, M. T., & Hixson, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review*, 33(2), 193–203. <https://doi.org/10.1080/02796015.2004.12086242>
- McKeown, M. G., Beck, I. L., Omanson, R. C., & Pople, M. T. (1985). Some effects of the nature and frequency of vocabulary instruction on the knowledge and use of words. *Reading Research Quarterly*, 20(5), 522–535. <https://doi.org/10.2307/747940>
- McGraw Hill. (2023). Wonders. <https://www.mheducation.com/prek-12/program/microsites/MKTSP-BGA10M0?srsId=AfmBOopEpdUScmFJCTAhqW7To8oqD8AbkQ12KDd1eADKWnPUa8rNLE6>
- Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, 23(3), 187–194. <https://doi.org/10.1177/01466219922031310>
- Minnesota Department of Education. (n.d.). *Minnesota Report Card*. https://rc.education.mn.gov/#assessmentsParticipation/orgId--999999000000_groupType--state_accountabilityFlg--FOC_NONE_test--MCA-III_subject--R_year--trend_grade--all_p--3
- Missall, K., Artman-Meeker, K., Roberts, C., & Ludeman, S. (2021). Implementing multitiered systems of support in preschool: Begin with universal screening. *Young Exceptional Children*, 24(4), 213–224. <https://doi.org/10.1177/1096250620931807>
- Missall, K. N., Hosp, M. K., & Hosp, J. L. (2019). Reading proficiency in elementary: Considering statewide testing, teacher ratings and rankings, and reading curriculum-

based measurement. *School Psychology Review*, 48(3), 267–275.

<https://doi.org/10.17105/SPR-2017-0152.V48-3>

Munger, K. A., & Murray, M. S. (2017). First-grade spelling scores within the dynamic indicators of basic early literacy skills (DIBELS) screening: An exploratory study. *Educational Assessment*, 22(2), 124–137.

<https://doi.org/10.1080/10627197.2017.1309275>

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, 20(2), 233–253. <https://doi.org/10.2307/747758>

National Academies of Sciences, Engineering, and Medicine. 2017. *Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress*. Washington, DC: The National Academies Press.

<https://doi.org/10.17226/23409>

National Assessment Governing Board. (2022). *Reading framework for the 2022 and 2024 National Assessment of Educational Progress*.

<https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/reading/2022-nagb-reading-framework-508.pdf>

National Center for Education Statistics. (n.d.). *The Nation's Report Card*.

<https://www.nationsreportcard.gov/highlights/reading/2022/>

National Center on Improving Literacy. (n.d.). *State of Dyslexia*. <https://www.stateofdyslexia.org>

National Center on Intensive Intervention. (n.d.). *iReady Diagnostic Reading*.

<https://charts.intensiveintervention.org/screening/tool/?id=4c5fbfb725845db3#DescriptiveInformation>

National Center on Intensive Intervention. (n.d.b.). *DIBELS (8th Edition) Oral Reading Fluency*.

<https://charts.intensiveintervention.org/screening/tool/?id=bf57b2df849f98fc>

National Center on Intensive Intervention. (2021). *Academic Screening Tools Chart*. U.S.

Department of Education, Office of Special Education Programs.

<https://charts.intensiveintervention.org/ascreening>

National Institute of Child Health and Human Development. (2000). *National Reading Panel:*

Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research

Literature on Reading and its Implications for Reading Instruction: Reports of the

Subgroups. National Institute of Child Health and Human Development, National

Institutes of Health.

<https://www.nichd.nih.gov/sites/default/files/publications/pubs/nrp/documents/report.pdf>

Nebraska Department of Education. (n.d.). *Nebraska Public School State Snapshot*.

<https://nep.education.ne.gov/#/profiles/state/full-profile/about?dataYears=20232024>

Nese, J. F. T., & Kamata, A. (2021). Addressing the large standard error of traditional CBM-R:

Estimating the conditional standard error of a model-based estimate of CBM-R.

Assessment for Effective Intervention, 47(1), 53–58.

<https://doi.org/10.1177/1534508420937801>

Nelson, P. M., Van Norman, E. R., & VanDerHeyden, A. (2017). Reduce, reuse, recycle: The

longitudinal value of local cut scores using state test data. *Journal of Psychoeducational*

Assessment, 35(7), 683–694. <https://doi.org/10.1177/0734282916658567>

New York State Education Department. (2024, August). *Testing Data to Inform Instructional*

Decisions and Individualized Student Learning Plans for the 2024-25 School Year.

<https://www.nysed.gov/news/2024/state-education-department-releases-preliminary-data-english-language-arts-mathematics-and-science-exams>

Ochs, S., Keller-Margulis, M. A., Santi, K. L., & Jones, J. H. (2020). Long-term validity and diagnostic accuracy of a reading computer-adaptive test. *Assessment for Effective Intervention*, 45(3), 210–225. <https://doi.org/10.1177/1534508418796271>

Office of Superintendent of Public Instruction (2025). *State Test Scores Show Continued Improvement in Math and English Language Arts*. <https://ospi.k12.wa.us/about-ospi/news-center/news-releases/state-test-scores-show-continued-improvement-math-and-english-language-arts>

Office of Superintendent of Public Instruction (n.d.). *Learning Assistance Program*. <https://ospi.k12.wa.us/student-success/support-programs/learning-assistance-program-lap>

Office of Superintendent of Public Instruction (n.d.b.). *State Testing Timelines and Calendar*. <https://ospi.k12.wa.us/student-success/testing/state-testing-timelines-calendar>

Office of Superintendent of Public Instruction (2023). *Learning Assistance Program Questions and Answers*. <https://ospi.k12.wa.us/sites/default/files/2023-08/lap-q-and.pdf>

Office of Superintendent of Public Instruction (2024). *Learning Assistance Program Guide*. <https://ospi.k12.wa.us/sites/default/files/2024-08/lap-guide.pdf>

Oslund, E. L., Elleman, A. M., & Wallace, K. (2021). Factors related to data-based decision-making: Examining experience, professional development, and the mediating effect of confidence on teacher graph literacy. *Journal of Learning Disabilities*, 54(4), 243–255. <https://doi.org/10.1177/0022219420972187>

Outhwaite, L. A., Aunio, P., Leung, J. K. Y., & Van Herwegen, J. (2024). Measuring mathematical skills in early childhood: A systematic review of the psychometric

- properties of early maths assessments and screeners. *Educational Psychology Review*, 36(4), 110-. <https://doi.org/10.1007/s10648-024-09950-6>
- Ozensoy, A. U. (2021). The effect of critical reading skill on academic success in social studies. *Eurasian Journal of Educational Research*, 21(93), 319-337. <https://doi.org/10.14689/ejer.2021.93.15>
- Ozernov-Palchik, O., Elizee, Z., Catania, F., Hacikamiloglu, M., Shattuck-Hufnagel, S., Petscher, Y., Ghosh, S., & Gabrieli, J. D. E. (2025). (Not so) universal literacy screening: a survey of educators reveals variability in implementation. *Annals of Dyslexia*, 75(3), 489–523. <https://doi.org/10.1007/s11881-025-00342-1>
- Paly, B. J., Klingbeil, D. A., Clemens, N. H., & Osman, D. J. (2022). A cost-effectiveness analysis of four approaches to universal screening for reading risk in upper elementary and middle school. *Journal of School Psychology*, 92, 246–264. <https://doi.org/10.1016/j.jsp.2022.03.009>
- Pearson, N. C. S. (2018). *AimswestPlus Introductory Guide*. NCS Pearson.
- Pedonti, S. F., Bratsch-Hines, M., McIntyre, N. S., & Aiken, H. (2024). Impacts of a teacher-delivered reading intervention on word-level fluency for early elementary students at risk for reading difficulties. *Reading & Writing Quarterly*, 1–19. <https://doi.org/10.1080/10573569.2024.2420943>
- Petscher, Y., & Kim, Y.-S. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology*, 49(1), 107–129. <https://doi.org/10.1016/j.jsp.2010.09.004>
- Polat, N., Zarecky-Hodge, A., & Schreiber, J. B. (2016). Academic growth trajectories of ELLs in NAEP data: The case of fourth- and eighth-grade ELLs and non-ELLs on mathematics

- and reading tests. *The Journal of Educational Research*, 109(5), 541–553.
<https://doi.org/10.1080/00220671.2014.993461>
- Purpura, D. J., Hume, L. E., Sims, D. M., & Lonigan, C. J. (2011). Early literacy and early numeracy: The value of including early literacy skills in the prediction of numeracy development. *Journal of Experimental Child Psychology*, 110(4), 647–658.
<https://doi.org/10.1016/j.jecp.2011.07.004>
- Quinn, J. M., Wagner, R. K., Petscher, Y., Roberts, G., Menzel, A. J., & Schatschneider, C. (2020). Differential co-development of vocabulary knowledge and reading comprehension for students with and without learning disabilities. *Journal of Educational Psychology*, 112(3), 608–627. <https://doi.org/10.1037/edu0000382>
- Rampey, B. D., Dion, G. S., & Donahue, P. L. (2009). *NAEP 2008: Trends in Academic Progress* (NCES 2009-479). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.
https://nces.ed.gov/nationsreportcard/pdf/main2008/2009479_1.pdf
- Reardon, S. F., Valentino, R. A., & Shores, K. A. (2012). Patterns of literacy among U.S. students. *The Future of Children*, 22(2), 17–37. <https://doi.org/10.1353/foc.2012.0015>
- Reed, D. K., & Sturges, K. M. (2013). An examination of assessment fidelity in the administration and interpretation of reading tests. *Remedial and Special Education*, 34(5), 259–268. <https://doi.org/10.1177/0741932512464580>
- Reed, D. K., Petscher, Y., & Truckenmiller, A. J. (2017). The contribution of general reading ability to science achievement. *Reading Research Quarterly*, 52(2), 253–266.
<https://doi.org/10.1002/rrq.158>

- Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the national assessment of educational progress (NAEP). *The American Psychologist*, 74(4), 445–458.
<https://doi.org/10.1037/amp0000356>
- Rice, M., Erbeli, F., Truckenmiller, A., Morris, J., & Coddling, R. S. (2024). Universal screening in kindergarten: Validity and classification accuracy of Istation's indicators of progress-early reading. *School Psychology*, 39(1), 72–80. <https://doi.org/10.1037/spq0000549>
- Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, 24(3), 309–338. <https://doi.org/10.1037/met0000184>
- Rindermann, H., & Thompson, J. (2013). Ability rise in NAEP and narrowing ethnic gaps? *Intelligence (Norwood)*, 41(6), 821–831. <https://doi.org/10.1016/j.intell.2013.06.016>
- Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology*, 46(3), 343–366.
<https://doi.org/10.1016/j.jsp.2007.06.006>
- Sabatini, J., Wang, Z., & O'Reilly, T. (2019). Relating reading comprehension to oral reading performance in the NAEP fourth-grade special study of oral reading. *Reading Research Quarterly*, 54(2), 253–271. <https://doi.org/10.1002/rrq.226>
- Sanchez, V. M., & O'Connor, R. E. (2022). Improving academic vocabulary for adolescent students with disabilities: A replication study. *Remedial and Special Education*, 43(2), 87–97. <https://doi.org/10.1177/07419325211016048>

- Sanders, E. A., Konold, T. R., & Cheng, Y. (2026). *Model-Based Effect Sizes for Multilevel Linear Regression Coefficients*. Manuscript under review at *Methodology*.
- Savage, R., Stuart, M., & Hill, V. (2001). The role of scaffolding errors in reading development: Evidence from a longitudinal and a correlational study. *British Journal of Educational Psychology*, 71(1), 1–13. <https://doi.org/10.1348/000709901158343>
- Schwanenflugel, P. J., Meisinger, E. B., Wisenbaker, J. M., Kuhn, M. R., Strauss, G. P., & Morris, R. D. (2006). Becoming a fluent and automatic reader in the early elementary school years. *Reading Research Quarterly*, 41(4), 496-522. <https://doi.org/10.1598/RRQ.41.4.4>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin. <http://catdir.loc.gov/catdir/enhancements/fy1105/2001131551-t.html>
- Shapiro, E. S., & Clemens, N. H. (2023). *Academic Skills Problems: Direct Assessment and Intervention* (5th Ed). Guilford.
- Shapiro, E. S., Dennis, M. S., & Fu, Q. (2015). Comparing computer adaptive and curriculum-based measures of math in progress monitoring. *School Psychology Quarterly*, 30(4), 470–487. <https://doi.org/10.1037/spq0000116>
- Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment*, 24(1), 19–35. <https://doi.org/10.1177/0734282905285237>

- Share, D. L., Jorm, A. F., Maclean, R., & Matthews, R. (1984). Sources of individual differences in reading acquisition. *Journal of Educational Psychology*, 76(6), 1309–1324.
<https://doi.org/10.1037/0022-0663.76.6.1309>
- Shhub, A., Jimenez, Z., & Solis, M. (2023). A synthesis of reading prosody: Evaluating phrasing and syntax interventions. *Reading & Writing Quarterly*, 39(6), 530–547.
<https://doi.org/10.1080/10573569.2022.2147464>
- Shin, J., & McMaster, K. (2019). Relations between CBM (oral reading and maze) and reading comprehension on state achievement tests: A meta-analysis. *Journal of School Psychology*, 73, 131–149. <https://doi.org/10.1016/j.jsp.2019.03.005>
- Shinn, M. R., Knutson, N., Good, R. H., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review*, 21(3), 459–479.
<https://doi.org/10.1080/02796015.1992.12085629>
- Shinn, M. R., Tindal, G. A., Spira, D., & Marston, D. (1987). Practice of learning disabilities as social policy. *Learning Disability Quarterly*, 10(1), 17–28.
<https://doi.org/10.2307/1510751>
- Silverman, R. D., McNeish, D., Speece, D. L., & Ritchey, K. D. (2021). Early screening for decoding- and language-related reading difficulties in first and third grades. *Assessment for Effective Intervention*, 46(2), 99–109. <https://doi.org/10.1177/1534508419857234>
- Smarter Balanced Assessment Consortium (2024). *Online Summative Test Administration Manual*. https://smarterbalanced.alohahsap.org/content/contentresources/en/Smarter-Balanced-Summative-TAM_2023-2024.pdf

Smarter Balanced Assessment Consortium (2017). *Estimated Testing Times for Smarter Balanced Summative Assessments*.

<https://portal.smarterbalanced.org/library/en/estimated-testing-times.pdf>

Smarter Balanced Assessment Consortium (n.d.). *A Decade of Advancing Teaching and Learning*. <https://smarterbalanced.org/our-system/>

Smarter Balanced Assessment Consortium (n.d.b.). *Sample Items*.

<https://sampleitems.smarterbalanced.org/BrowseItems?Claim=ELA2&Subject=ELA&Grade=1>

Smarter Balanced Assessment Consortium (2012). *Claims for English Language Arts/Literacy Summative Assessment*. <https://portal.smarterbalanced.org/library/en/claims-for-the-ela-literacy-summative-assessments.pdf>

Smarter Balanced Assessment Consortium (2013). *Initial Achievement Level Descriptors and College Content-Readiness Policy*.

<https://portal.smarterbalanced.org/library/en/elaliteracy-alds-and-college-content-readiness-policy.pdf>

Smarter Balanced Assessment Consortium (2022). *2020-2021 Summative Technical Report*.

https://technicalreports.smarterbalanced.org/2020-21_summative-report/_book/index.html

Smart, D., Youssef, G. J., Sanson, A., Prior, M., Toumbourou, J. W., & Olsson, C. A. (2017).

Consequences of childhood reading difficulties and behaviour problems for educational achievement and employment in early adulthood. *British Journal of Educational*

Psychology, 87(2), 288–308. <https://doi.org/10.1111/bjep.12150>

- Spear-Swerling, L., & Cheesman, E. (2012). Teachers' knowledge base for implementing response-to-intervention models in reading. *Reading & Writing, 25*(7), 1691–1723. <https://doi.org/10.1007/s11145-011-9338-3>
- Spees, L. P., Potochnick, S., & Perreira, K. M. (2016). The academic achievement of limited English proficient (LEP) youth in new and established immigrant states: Lessons from the national assessment of educational progress (NAEP). *Education Policy Analysis Archives, 24*(99), 99. <https://doi.org/10.14507/epaa.24.2130>
- Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*(3), 407–419. <https://doi.org/10.1080/02796015.2001.12086123>
- Stanovich, K. E., Cunningham, A. E., & Feeman, D. J. (1984). Relation between early reading acquisition and word decoding with and without context: A longitudinal study of first-grade children. *Journal of Educational Psychology, 76*(4), 668–677. <https://doi.org/10.1037/0022-0663.76.4.668>
- Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42*(8), 795–819.
- Sutter, C. C., Campbell, L. O., & Lambie, G. W. (2020). Predicting second-grade students' yearly standardized reading achievement using a computer-adaptive assessment. *Computers in the Schools, 37*(1), 40–54. <https://doi.org/10.1080/07380569.2020.1720611>
- Texas Education Agency. (2004). *Student Assessment Division Technical Digest 2004–2005*. <http://ritter.tea.state.tx.us/student.assessment/resources/techdig05/index.html>

- Thomas, A. S., & January, S.-A. A. (2021). Evaluating the criterion validity and classification accuracy of universal screening measures in reading. *Assessment for Effective Intervention*, 46(2), 110–120. <https://doi.org/10.1177/1534508419857232>
- Torgesen, J., Wagner, R., & Rashotte, C. (1999). *Test of Word Reading Efficiency*. Pro-Ed.
- Tortorelli, L. S., Strong, J. Z., & Anderson, B. E. (2024). Multisyllabic decoding achievement and relation to vocabulary at the end of elementary school. *Journal of Experimental Child Psychology*, 246, 106018-. <https://doi.org/10.1016/j.jecp.2024.106018>
- University of Houston. (2008). *Texas Middle School Fluency Assessment*. University of Houston.
- University of Oregon (n.d.). *8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®): Materials*. <https://dibels.uoregon.edu/materials>
- University of Oregon (n.d.b). *8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®): Frequently Asked Questions*. <https://dibels.uoregon.edu/faqs#r150>
- University of Oregon (2020). *8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®): Technical Manual*. <https://dibels.uoregon.edu>
- University of Oregon (2023a). *DIBELS® 8th Edition Benchmark Goals*. https://dibels.uoregon.edu/sites/default/files/2024-01/dibels8_benchmark_goals.pdf
- University of Oregon. (2023b). *8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Administration and Scoring Guide (8th ed.)*. <https://dibels.uoregon.edu>
- University of Oregon, Center on Teaching and Learning (2018). *Understanding the Research behind DIBELS® 8th Edition*. Technical Report 1801. https://dibels.uoregon.edu/sites/default/files/DIBELS8thEdition_TechRpt1801_ResearchBrief.pdf

- Van Norman, E. R., & Forcht, E. R. (2024). Modeling growth in reading with curriculum-based measurement versus computer-adaptive tests. *Assessment for Effective Intervention*, 50(1), 29–39. <https://doi.org/10.1177/15345084241292731>
- Van Norman, E. R., Nelson, P. M., & Klingbeil, D. A. (2017). Single measure and gated screening approaches for identifying students at-risk for academic problems: Implications for sensitivity and specificity. *School Psychology Quarterly*, 32(3), 405–413. <https://doi.org/10.1037/spq0000177>
- Van Norman, E. R., Nelson, P. M., & Parker, D. C. (2018). Curriculum-based measurement of reading decision rules: Strategies to improve the accuracy of treatment recommendations. *School Psychology Review*, 47(4), 333–344. <https://doi.org/10.17105/SPR-2017-0089.V47-4>
- VanDerHeyden, A. M., Burns, M. K., Bonifay, W., & Eckert, T. (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>
- VanDerHeyden, A. M., Coddling, R. S., & Martin, R. (2017). Relative value of common screening measures in mathematics. *School Psychology Review*, 46(1), 65–87. <https://doi.org/10.1080/02796015.2017.12087608>
- VanMeveren, K., Hulac, D., & Wollersheim-Shervey, S. (2020). Universal screening methods and models: Diagnostic accuracy of reading assessments. *Assessment for Effective Intervention*, 45(4), 255–265. <https://doi.org/10.1177/1534508418819797>
- Vanneman, A., Hamilton, L., Anderson, J. B., & Rahman, T. (2009). *Achievement Gaps: How Black and White Students in Public Schools Perform in Mathematics and Reading on the*

- National Assessment of Educational Progress*, (NCES 2009-455). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. <https://files.eric.ed.gov/fulltext/ED505903.pdf>
- Vaughn, S., Fletcher, J. M., Francis, D. J., Denton, C. A., Wanzek, J., Wexler, J., Cirino, P. T., Barth, A. E., & Romain, M. A. (2008). Response to intervention with older students with reading difficulties. *Learning and Individual Differences*, 18(3), 338–345. <https://doi.org/10.1016/j.lindif.2008.05.001>
- Vaughn, S., & Wanzek, J. (2024). Promoting adolescents' comprehension of text: Efficacy and effectiveness. *Remedial and Special Education*, 45(1), 58–67. <https://doi.org/10.1177/07419325231190805>
- Vilenius-Tuohimaa, P. M., Aunola, K., & Nurmi, J.-E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology*, 28(4), 409–426. <https://doi.org/10.1080/01443410701708228>
- Wanzek, J., Martinez, L., Fall, A.-M., Roberts, G., Stillman, S., & Kent, S. C. (2018). Text reading supports in social studies content instruction and their relationship to student knowledge acquisition. *Reading & Writing Quarterly*, 34(4), 349–360. <https://doi.org/10.1080/10573569.2018.1446858>
- Washburn, J. (2022). Reviewing evidence on the relations between oral reading fluency and reading comprehension for adolescents. *Journal of Learning Disabilities*, 55(1), 22-42. <https://doi.org/10.1177/00222194211045122>
- Wayman, M.M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education*, 41(2), 85–120. <https://doi.org/10.1177/00224669070410020401>

- Weisenburgh-Snyder, A. B., Malmquist, S. K., Robbins, J. K., & Lipshin, A. M. (2015). A model of MTSS: Integrating precision teaching of mathematics and a multi-level assessment system in a generative classroom. *Learning Disabilities: A Contemporary Journal*, 13(1), 21-41.
- Wells, C. S., Baldwin, S., Hambleton, R. K., Sireci, S. G., Karatonis, A., & Jirka, S. (2009). Evaluating score equity assessment for state NAEP. *Applied Measurement in Education*, 22(4), 394–408. <https://doi.org/10.1080/08957340903221683>
- Wolters, A. P., Kim, Y.-S. G., & Szura, J. W. (2022). Is reading prosody related to reading comprehension? A Meta-analysis. *Scientific Studies of Reading*, 26(1), 1–20. <https://doi.org/10.1080/10888438.2020.1850733>
- White, S., Sabatini, J., Park, B. J., Chen, J., Bernstein, J., & Li, M. (2021). The 2018 NAEP oral reading fluency study (NCES 2021–025). *National Center for Education Statistics*. <https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2021025>.
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.
- 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>
- Zhu, Y. (2022). Reading matters more than mathematics in science learning: an analysis of the relationship between student achievement in reading, mathematics, and science. *International Journal of Science Education*, 44(1), 1–17. <https://doi.org/10.1080/09500693.2021.2007552>

Table 1*Sample Demographic Characteristics*

Characteristic	Grade 4 (<i>n</i> = 615)		Grade 5 (<i>n</i> = 576)		Combined (<i>n</i> = 1,191)	
	<i>n</i>	(%)	<i>n</i>	(%)	<i>n</i>	(%)
<i>Gender</i>						
Female	287	(47%)	279	(48%)	566	(48%)
Male	328	(53%)	296	(51%)	624	(52%)
Non-Binary	0	(0%)	1	(<1%)	1	(<1%)
<i>Race</i>						
Asian	11	(18%)	9	(2%)	20	(2%)
Black	3	(<1%)	2	(<1%)	5	(<1%)
Latine	42	(7%)	62	(11%)	104	(9%)
Native American/ Hawaiian/ Pacific Islander	2	(<1%)	4	(<1%)	6	(1%)
Two or More Races	57	(9%)	53	(9%)	110	(9%)
White	500	(81%)	446	(77%)	946	(79%)
<i>Multilingual Services</i>						
Yes	15	(2%)	18	(3%)	33	(3%)
No	600	(98%)	558	(97%)	1158	(97%)
<i>Special Education Services</i>						
Yes	94	(15%)	103	(18%)	197	(17%)
No	521	(85%)	473	(82%)	994	(83%)
<i>LAP Services</i>						
Yes	68	(11%)	54	(9%)	122	(10%)
No	547	(89%)	522	(91%)	1069	(90%)

Note. LAP = Learning Assistance Program.

Table 2*SBA-ELA Reading Levels and Skill Expectations for Students in Grades 4-5*

SBA Level	Text Complexity	Grade 4 Score Range	Grade 4 Skills	Grade 5 Score Range	Grade 5 Skills
1	Low	2140 – 2415	Can find information from the text that minimally supports inferences that are provided	2200 – 2441	Can use details from a passage to make inferences when reading texts of low complexity
2	Moderate	2416 – 2472	Can find information from text that partially support inferences	2442 – 2501	Can use details from texts to make inferences when reading passages of moderate complexity
3	Moderate-High	2473 – 2532	Can find textual evidence that supports inferences when reading texts of moderate-to-high complexity	2502 – 2581	Can identify implicit information that supports inferences for texts of moderate-to-high complexity
4	High	2533 – 2650+	Can identify and explain the textual evidence that supports inferences when reading unusually challenging texts	2582 – 2700+	Can identify implicit information to make inferences about texts that are unusually challenging

Note. SBA Level 3 = Meets State Expectations.

Table 3*Sample Descriptive Statistics and Zero-Order Correlations*

Variable	<i>M</i>	(<i>SD</i>)	1.	2.	3.	4.	5.	6.	7.	8.	9.
<i>Dependent Variable</i>											
1. SBA-ELA Spring 2024	2531.81	(96.95)	--								
<i>Reading Screener (Student Level)</i>											
2. SBA-ELA 2023 Student (CMC)	0.00	(94.52)	.74	--							
3. iReady Reading Student (CMC)	0.00	(48.99)	.77	.81	--						
4. DIBELS-ORF Student (CMC)	0.00	(38.02)	.60	.65	.71	--					
<i>Reading Screeners (School Level)</i>											
5. SBA-ELA 2023 School Mean	2484.22	(23.14)	.20	.00	.00	.00	--				
6. iReady Reading School Mean	567.29	(10.91)	.19	.00	.00	.00	.93	--			
7. DIBELS-ORF School Mean	115.13	(9.80)	.18	.00	.00	.00	.91	.94	--		
<i>Student-Level Covariates</i>											
8. Grade Level (1 = Gr. 5, 0 = Gr. 4)	0.48	(0.50)	.22	.26	.25	.14	.01	.01	.01	--	
9. LAP Status (1 = yes, 0 = no)	0.10	(0.30)	-.32	-.32	-.33	-.42	-.08	-.06	-.03	-.03	--

Note. *N* = 1191 students (L1) from 10 schools (L2). SBA-ELA = Smarter Balanced Assessment-English Language Arts; iReady Reading = iReady Reading Diagnostic; DIBELS-ORF = Dynamic Indicators of Basic Early Literacy Skills-Oral Reading Fluency subtest. CMC = cluster-mean-centered (student score – school mean). Reading Screeners include SBA-ELA measured in spring 2023, iReady Reading measured in fall 2023, and DIBELS-ORF measured in fall 2023. LAP = Learning Assistance Program. Pearson's *r* correlations reported; values significant at the .05, 2-tailed level are boldfaced.

Table 4*Predicting Spring 2024 SBA-ELA with Each Reading Screener in Isolation (Model 1)*

<i>Fixed Effects</i>	Reading Screener: Prior Year SBA-ELA				Reading Screener: Fall iReady Reading				Reading Screener: Fall DIBELS-ORF			
	<i>Coeff</i>	<i>(SE)</i>	<i>p</i>	<i>sr²</i>	<i>Coeff</i>	<i>(SE)</i>	<i>P</i>	<i>sr²</i>	<i>Coeff</i>	<i>(SE)</i>	<i>p</i>	<i>sr²</i>
Mean SBA-ELA (Intercept)	2528.03	(7.42)	<.001	--	2529.00	(7.34)	<.001	--	2528.73	(7.54)	<.001	--
Grade (1 = Gr 5, -1 = Gr 4)	2.71	(1.97)	.169	0.1%	3.11	(1.85)	.093	0.1%	10.48	(2.67)	<.001	0.7%
LAP Status (1 = yes, -1 = no)	-14.93	(4.41)	.001	0.4%	-7.20	(4.21)	.088	0.1%	-8.87	(8.64)	.305	0.1%
Grade x LAP	-2.57	(4.39)	.558	0.0%	-7.35	(4.16)	.077	0.1%	-20.90	(8.49)	.014	0.3%
Reading Screener (CMC, Z)	68.53	(1.97)	<.001	40.2%	72.17	(1.87)	<.001	44.2%	54.21	(2.55)	<.001	21.5%
Grade x Reading Screener	-1.69	(1.98)	.393	0.0%	-0.39	(1.89)	.836	0.0%	-3.13	(2.57)	.223	0.1%
LAP x Reading Screener	-4.44	(3.62)	.220	0.1%	2.24	(3.65)	.540	0.0%	-0.69	(6.59)	.917	0.0%
Grade x LAP x Reading Screener	-1.93	(3.68)	.600	0.0%	-0.32	(3.73)	.931	0.0%	-12.05	(6.55)	.066	0.2%
<i>Random Effects</i>	<i>Var</i>				<i>Var</i>				<i>Var</i>			
Between Schools (Intercepts)	22.57				22.38				22.15			
Within Schools (Students)	61.51				58.05				73.06			
<i>Model Fit</i>	<i>Value</i>				<i>Value</i>				<i>Value</i>			
Fixed Effects R^2	54.3%				58.7%				38.2%			
Akaike Information Criterion (AIC)	13240				13103				13646			
Bayesian Information Criterion (BIC)	13291				13154				13697			
Deviance (-2*LogLikelihood)	13220				13083				13626			
Model Residual df	1181				1181				1181			

Note. $N = 1191$ students (L1) within 10 schools (L2). All predictors were cluster-mean centered with categorical predictors effect-coded and continuous predictors standardized. Models estimated using full maximum likelihood by R lme4; t -test p -values based on Satterthwaite df computed using lmerTest. sr^2 = unique variance explained, based on total fixed effects R^2 . Observed p -values reported; values in boldface indicate significance at .05, 2-tailed.

Table 5*Predicting Spring 2024 SBA-ELA with Combinations of Reading Screeners (Model 2)*

<i>Fixed Effects</i>	Model 2a:				Model 2b:				Model 2c:			
	SBA-ELA + iReady Reading				SBA-ELA + DIBELS-ORF				All 3 Screeners			
	<i>Coeff</i>	<i>(SE)</i>	<i>p</i>	<i>sr</i> ²	<i>Coeff</i>	<i>(SE)</i>	<i>p</i>	<i>sr</i> ²	<i>Coeff</i>	<i>(SE)</i>	<i>p</i>	<i>sr</i> ²
Mean SBA-ELA (Intercept)	2529.34	(7.21)	<.001	--	2529.93	(7.41)	<.001	--	2529.63	(7.28)	<.001	--
Grade (1 = Gr 5, -1 = Gr 4)	0.89	(1.79)	.617	0.0%	1.30	(2.23)	.560	0.0%	0.53	(2.06)	.797	0.0%
LAP Status (1 = yes, -1 = no)	-2.83	(4.20)	.500	0.0%	1.46	(7.16)	.838	0.0%	0.12	(6.60)	.986	0.0%
Grade x LAP	-6.06	(4.15)	.145	0.1%	-14.35	(7.04)	.042	0.1%	-9.40	(6.50)	.148	0.1%
Prior SBA-ELA (CMC, Z)	32.20	(2.81)	<.001	3.5%	57.11	(2.35)	<.001	18.5%	31.05	(2.83)	<.001	3.2%
Grade x Prior SBA-ELA	3.23	(2.81)	.251	0.0%	-0.69	(2.36)	.770	0.0%	2.65	(2.84)	.352	0.0%
LAP x Prior SBA-ELA	-3.57	(4.35)	.412	0.0%	-0.98	(3.76)	.794	0.0%	-2.96	(4.35)	.496	0.0%
Grade x LAP x Prior SBA-ELA	0.73	(4.38)	.868	0.0%	-0.21	(3.81)	.957	0.0%	1.08	(4.38)	.805	0.0%
iReady Reading (CMC, Z)	46.94	(2.82)	<.001	7.4%	--	--	--	--	43.70	(3.06)	<.001	5.4%
Grade x iReady Reading	-4.39	(2.83)	.121	0.1%	--	--	--	--	-4.42	(3.07)	.150	0.1%
LAP x iReady Reading	6.17	(4.63)	.183	0.1%	--	--	--	--	7.07	(4.86)	.147	0.1%
Grade x LAP x iReady Reading	-1.53	(4.69)	.744	0.0%	--	--	--	--	-0.60	(4.94)	.903	0.0%
DIBELS-ORF (CMC, Z)	--	--	--	--	20.75	(2.50)	<.001	2.2%	6.40	(2.50)	.011	0.2%
Grade x DIBELS-ORF	--	--	--	--	-3.89	(2.52)	.123	0.1%	0.16	(2.51)	.949	0.0%
LAP x DIBELS-ORF	--	--	--	--	3.63	(5.72)	.526	0.0%	-0.58	(5.57)	.917	0.0%
Grade x LAP x DIBELS-ORF	--	--	--	--	-8.50	(5.65)	.133	0.1%	-3.58	(5.51)	.516	0.0%
<i>Random Effects</i>	<i>Var</i>				<i>Var</i>				<i>Var</i>			
Between Schools (Intercepts)	22.03				22.23				21.97			
Within Schools (Students)	55.02				59.64				54.85			
<i>Model Fit</i>	<i>Value</i>				<i>Value</i>				<i>Value</i>			
Fixed Effects <i>R</i> ²	62.3%				56.7%				62.5%			
Akaike Information Criterion (AIC)	12984				13175				12985			
Bayesian Information Criterion (BIC)	13055				13246				13076			
Deviance (-2*LogLikelihood)	12956				13147				12949			
Model Residual <i>df</i>	1177				1177				1173			

Note. *N* = 1191 students (L1) within 10 schools (L2). All predictors were cluster-mean centered with categorical predictors effect-coded and continuous predictors standardized. Models estimated using full maximum likelihood by *R* lme4; *t*-test *p*-values based on Satterthwaite *df* computed using lmerTest. *sr*² = unique variance explained, based on total fixed effects *R*². Observed *p*-values reported; values in boldface indicate significance at .05, 2-tailed.

Figure 1

Model-Predicted Mean Relations between Reading Screeners and Spring 2024 SBA-ELA (Model 1)

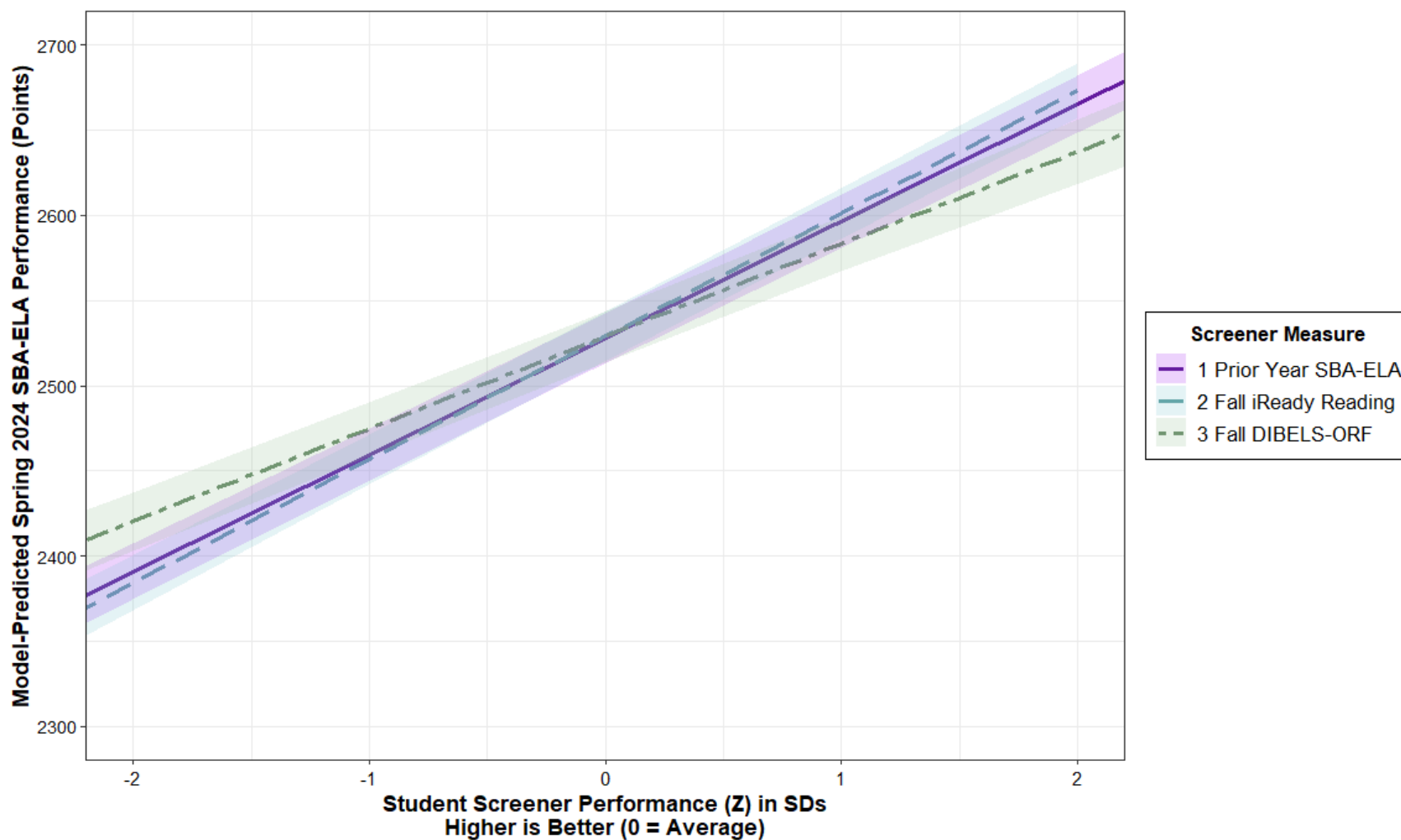


Figure 2

Model-Predicted Relations between Reading Screeners and Spring 2024 SBA-ELA by Grade and LAP Status (Model 1)

