THE DIAGNOSTIC ACCURACY OF THREE COMPUTER-ADAPTIVE SCREENING MEASURES OF READING

by

Susan Barnes Porter

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

Middle Tennessee State University

December 2022

Dissertation Committee:

Dr. Timothy Odegard, Chair

Dr. Amy Elleman

Dr. Eric Oslund

Dr. Emily Farris

ACKNOWLEDGEMENTS

This work would not have been possible without the financial support of the Literacy Studies Ph.D. program throughout my degree and the gracious support of numerous colleagues, family, and friends.

I would like to express my sincere gratitude to my advisor, Dr. Timothy Odegard, and my dissertation committee, Dr. Amy Elleman, Dr. Eric Oslund, and Dr. Emily Farris for all the guidance and support given to me throughout my program. Your motivation and immense knowledge have been invaluable to me during my pursuit of this degree. Thank you for your enduring patience, sage advice, and commitment to my success.

To my amazing cohort – Jessica Dainty, Jennifer Grow, Molly Risley, and Bingshi Zhang – I want to say thank you for the friendship, the encouragement, and all the laughs. This journey was made all the better for having you walk alongside me. I am truly blessed to count you as my friends.

I am beyond blessed to have the love and encouragement of my amazing husband, Spencer, without whom I could not have achieved this dream. Thank you for believing in me and cheering me on every step of the way.

To my wonderful children, Lake, Sarah Kate, and Savannah – thank you for inspiring me to continue to pursue my dreams. To my mom and dad - your love and belief in me have meant so much to me. You are all truly a blessing to me.

ABSTRACT

The data from universal screeners must be valid and reliable in order to use it to make appropriate decisions about how best to allocate resources to support students who are at risk of not passing the state achievement test. The instruments used as part of universal screening must also have diagnostic accuracy. This study examined the diagnostic accuracy of three computer-adaptive universal screening measures of reading for predicting student performance on the state achievement test. Using extant data, spring universal screener scores from second graders in public schools in a southeastern state were used to predict their performance on the state achievement test administered in the spring of their third grade year. Logistic regression and receiver operating characteristic curves were used to examine the diagnostic accuracy of Istation Indicators of Progress Early Reading, MAP-Reading, and Star Reading. Results show that all three computer-adaptive universal screeners were strong predictors of the ACT Aspire, an assessment used as the statewide achievement test. However, the Istation and MAP universal screeners were reliably more accurate than the Star Reading assessment based on area under the curve comparisons. In analyses of based on cut scores associated with various means of standard setting, none of the screeners met the recommended levels of .90 sensitivity and .70 for specificity. Implications for use of the computer-adaptive screeners in schools as part of a universal screening process are discussed.
Keywords: computer-adaptive test, universal screening, achievement measures, diagnostic accuracy

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Learning to read proficiently in the early years of formal schooling is a foundational skill for future academic achievement and is essential for successful life outcomes. Early reading development is a strong predictor of later reading development and academic success (Jeon et al., 2018; Juel, 1988). Therefore, early identification of students at risk for academic difficulty is essential in supporting future academic achievement and preventing adverse life outcomes associated with academic failure, particularly reading failure.

The challenges associated with poor reading skills are well-documented (Lyon, 1998; Snow et al., 1998). Students with low reading skills are four times more likely to drop out of high school (Anne E. Casey Foundation, 2010). The job and earning prospects for high school dropouts are limited. In a recent study, researchers found that adults with a minimum proficiency level of literacy (Level 3) had a significantly higher average annual income of $63,000 compared to adults just below proficiency (Level 2) with an average annual income of $48,000. Adults with the lowest level of literacy (Level 1) earned just $34,000 annually (Nietzel, 2020). High school dropouts are also more likely to become incarcerated (Smart et al., 2017). Morken and colleagues (2021) reviewed 20 years of research on the prevalence of reading disabilities in prison populations and found that prevalence rates are much higher than in the general population, with up to 50% of the prison population having reading disorders. Knowing the harsh reality of outcomes for individuals with poor reading skills, educators must have the necessary tools to identify students at risk as early as possible so resources can be allocated to intervene and prevent such dire outcomes (Lyon et al., 2005).

Assessment in Schools

One indicator of academic proficiency or lack thereof is performance on state achievement measures. Standardized assessments, which are uniform assessments given to an entire population of students, are one type of common assessment that has been around for more than a century (Hutt & Schneider, 2018). Statewide assessments are considered a primary predictor of future achievement and an indicator of education effectiveness (Van Norman et al., 2017). First appearing in 1845, standardized testing slowly replaced oral recitation and public exhibition practices to measure how well students were learning and how effective the local schoolmaster was (Reese, 2013). Applying statistics to education introduced a new way to perceive education. At that time, assessment was considered to be in the domain of psychometricians, who were considered experts in quantifying academic learning despite their lack of understanding of schooling. In fact, after World War I, large districts began creating departments dedicated to research and measurement staffed with these experts whose job was to track students efficiently and systematically (Hutt & Schneider, 2018). Despite the control shifting from teachers to the commercial enterprise of testing, the data collected by these departments did allow school leaders to make more informed decisions and draft policies based on data (Hutt & Schneider, 2018). It also provided a means to track student growth yearly and compare students' scores (Hutt & Schneider, 2018). These logical reasons for using assessments became embedded in education. For a society that valued science and the scientific approach to the world, education during the early 20th century provided a new field to apply science. By 1920, standardized assessments had become an accepted educational practice.

In 1965, the passage of the Elementary and Secondary Education Act (ESEA) created the first federal provision for testing, which led to the requirement of state testing for some students. As part of President Lyndon B. Johnson's "War on Poverty" campaign, the ESEA created Title I, a program designed to provide additional funding to districts and schools that served a high population of students from low-income families (Paul, 2016). The additional funding came with higher accountability standards which required assessing whether or not students' skill gaps were closing. Although only certain students fell into the required testing category, many schools opted to test all students for expediency (Hutt & Schneider, 2018).

In 2001, the No Child Left Behind (NCLB) legislation was passed, which increased the accountability of schools and districts. This law mandated measures of adequate yearly progress as measured by statewide standardized tests and created what some consider the most expansive testing requirements in history (Hutt & Schneider, 2018; Shapiro et al., 2006). It also tied educational funding to testing outcomes which many argue led to a narrowed curriculum focused primarily on passing the state tests (Hutt & Schneider, 2018). In 2015, NCLB was replaced with the Every Student Succeeds Act (ESSA). This new legislation addressed how much testing is necessary, not if students should be tested (Hutt & Schneider, 2018). It still required statewide standardized testing and reporting those results to educational stakeholders.

In addition to federal accountability legislation, several contextual factors have also contributed to the continuation of testing in our educational system. First, the decentralized nature of our educational system means that individual states have the responsibility of drafting educational standards and choosing instructional materials for

public schooling in the state. Therefore, many argue that the eclecticism of the educational system requires common measures to ensure a quality educational program from one state to the next (Hutt & Schneider, 2018). Common measures also allow colleges to compare students from various school settings when making admission decisions.

Another factor that may have led to the widespread reliance on testing is the belief that schools should find and reward talent. This belief, known as a meritocracy, dates back to before the founding of the United States of America (Plato, 1968). Testing became the means through which talent and aptitude could be identified. Along with this belief was the idea that testing would motivate students to do well in school so that their talents could be recognized and rewarded. This concept of meritocracy is closely linked to the belief that success in school leads to success in life.

A third contextual factor leading to the continued focus on testing in schools is the funding structure for public education. Schools are often funded through tax-supported systems (Hutt & Schneider, 2018). Testing is one means by which schools and districts are held accountable for the proper use of those funds. External stakeholders such as taxpayers want to know that their taxes are being used properly and that they are getting a return on their investment. Assessments provided a measure of how effectively those dollars are being used.

These contextual factors and the federal accountability legislation have perpetuated the reliance on and persistence of state testing programs in American education. Despite an initial focus on accountability at the district and school levels, testing programs have increasingly pushed the focus to the individual student and how to

raise individual student's scores (Silberglitt & Hintze, 2005). This focus on individual students has also led to higher stakes in testing performance (Sutter et al., 2020).

Although statewide standardized tests are required annually under federal legislation, the design and format of the measures have been left to the individual states. State departments of education can use a commercial test created by a testing development company or a test developed by the state (Hutt & Schneider, 2018). While commercially developed tests are considered more psychometrically sound, state-developed tests are highly correlated with state standards since these measures are often written for a single state (Shin & McMaster, 2019).

The persistence of testing in American education programs has come with critiques. Some critics have argued that testing distorts the education process by narrowing the curriculum to focus primarily on what is being tested (Hutt & Schneider, 2018). Others believe that the time and money spent on testing programs is wasted since it often just confirms what is already known about the state of the current system. Still, others believe that testing is used to erroneously classify and compare schools, teachers, and students when the conditions and settings are not equal. Yet, these critiques have led to more testing, not less (Hutt & Schneider, 2018).

Nevertheless, when used with intended targets and appropriateness, standardized testing measures can provide important educational information. End of year assessments may be the most reliable information about a student's progress. It is reasonable to expect that students demonstrate proficiency with reading and understanding grade level texts at the end of Grade 3 because they have had three years of formal schooling before entering the tested grade (Schatschneider et al., 2008). Achievement tests can also support

identifying specific areas of academic weakness (Francis et al., 2005). However, waiting until the end of Grade 3 to determine if students are on track for success is too late (Silberglitt & Hintze, 2005).

Universal Screening Assessments

Screening is one tool available to educators for early identification of students who are at risk for not passing the state test. The concept of academic screening for reading risk is based on medical screenings, a key strategy in preventive medicine (Armstrong & Eborall, 2012). In medicine, screening is offered to all people within an age or sex group to identify people likely to have or develop a particular disease or condition (Armstrong & Eborall, 2012). The screening results are then connected to additional diagnostics to determine if there is a need for a treatment plan to be developed. In addition, screening can identify the risk of a future health issue allowing for preventive measures to be identified and implemented to decrease the likelihood that a person will develop the disease. In education, screening is a proactive means of identifying students at risk for reading difficulty and providing them with appropriate preventive instruction and intervention (Fuchs & Fuchs, 2009).

Identifying students at risk for reading difficulties in the early grades is critical for supporting future academic achievement. Traditionally, students had to show significant difficulty before assessments were completed (Vaughn & Fuchs, 2006). The reauthorization of the Individuals with Disabilities Education Act (2004) sought to address this “wait to fail” model by including a provision for a process that consisted of early screening and multiple tiers of increasingly intensive interventions (Fuchs & Fuchs, 2009). This process, known as response to intervention (RTI), shifted the focus from

"wait to fail" to early identification and intervention designed to prevent reading difficulties (Al Otaiba et al., 2019).

Early identification and intervention are motivated by the established consensus that the earlier the intervention, the better the outcomes for students (Fletcher et al., 2019). Identifying students at risk is a critical first step in supporting later academic achievement (Fletcher et al., 2021; Vaughn & Fuchs, 2006). Identifying risk for reading difficulties can occur as early as kindergarten and first grade before reading difficulty begins (Catts et al., 2015). In fact, assessments of children's development at the age of five have predicted academic skills at age ten (Jeon et al., 2018). Without early identification and intervention, students will continue to fall further behind their peers throughout their schooling and increase their likelihood of dropping out of school (Sutter et al., 2020). Identifying students at risk for future academic difficulties and determining how to accelerate their growth is a complex issue that requires a consistent and accurate process for identification (Al Otaiba & Petscher, 2020; Hosp et al., 2011).

Currently, a majority of states have a multi-tiered approach to instruction and intervention (Bailey, 2019). The most common frameworks for this approach are Response to Intervention (RTI) and Multi-Tiered System of Supports (MTSS) (Leonard et al., 2019). Although both frameworks incorporate increasingly intensive tiers of instruction and intervention for students identified as at-risk or struggling, there are nuanced differences between the two frameworks (Bailey, 2019, Oslund et al., 2021). For this study, we will use RTI to refer to a multi-tiered approach to instruction and intervention for consistency.

Despite the widespread use of a multi-tiered approach to instruction and intervention, many schools have failed to implement it with the fidelity necessary to ensure efficacy (Al Otaiba et al., 2019; Fuchs & Fuchs, 2017). Teachers reported having a broad understanding of RTI and its assessments but lack the knowledge on how to use the data to make informed decisions (Al Otaiba et al., 2019). They also reported misunderstanding the data-based approach of RTI, often thinking the process is only for determining special education eligibility or that it is just for struggling readers. However, the process is intended to provide differentiated instruction and experiences for all students (Whittaker & Batsche, 2019).

Assessment is an essential first step in any RTI framework. Assessment is the process of measuring the characteristics of objects or people to obtain data or information (Hosp & Ardoin, 2008). In RTI, the initial assessment given is a universal screener, an assessment given to all students in a grade, school, or district (Hosp & Ardoin, 2008). The data or information gained from the universal screener is then used to decide on appropriate intervention placements (Lam & McMaster, 2014). Identification must be carried out in a timely and accurate manner so that the interventions selected can be of the appropriate intensity and individualized for students who demonstrate risk (Catts et al., 2015; Compton et al., 2012). Therefore, the assessments must provide opportunities for the students to perform the skills on which decisions are based (Hosp & Ardoin, 2008).

Universal screeners can be standards-based or skills-based, administered to a whole group or in one-on-one settings, and be print-based or computer-based. Ideally, screeners are inexpensive, brief, easy to administer and interpret, and linked to instruction

(Petscher et al., 2011). With the range of measures available as universal screeners, establishing the predictive validity and diagnostic accuracy of the measures is crucial (Ball & O'Connor, 2016). Predictive validity indicates how well student performance on one measure predicts performance on another measure (Hosp et al., 2011). Diagnostic accuracy indicates how accurately a measure categorizes student performance using dichotomous or continuous predictors and dichotomous outcomes (e.g., pass/fail) (Ball & O'Connor, 2016; Klingbeil et al., 2018). To be effective, universal screeners should be able to differentiate students who are at risk for reading difficulties from students who are not at risk while also limiting the number of false positives and false negatives. Due to the impact of testing on instructional time, universal screeners should have high levels of diagnostic accuracy, allowing them to predict if a student will pass or fail high-stakes measures, such as a state achievement test as a single measure. Valuable time and resources can be wasted if screeners do not meet this need (Klingbeil et al., 2015; Lam & McMaster, 2014).

Therefore, the process for selecting a universal screener should include consideration of the gold standard measure (e.g., state achievement test) the screener is predicting and the psychometric properties that maximize the diagnostic accuracy of the measure, both of which can have implications for the allocation of resources in a school (Kent et al., 2019; Petscher et al., 2011). The gold standard measure selected defines the risk being predicted (Petscher et al., 2019b). For most schools, the year-end state achievement test is a measure on which their performance is judged and, therefore, a measure of great importance in the elementary grades (Kent et al., 2019). Hence, having a universal screener that can accurately identify students at risk of not passing the year-end

state test has considerable practical value. This study compares the diagnostic accuracy of three computer-adaptive tests used as universal screeners to predict the risk of not passing a state achievement test.

CHAPTER II: REVIEW OF LITERATURE

Assessment has long been a central practice in schooling. Both formative and summative measures of student learning have taken many forms and served various purposes throughout the history of education (Hutt & Schneider, 2018). At the same time, testing programs have also faced staunch criticism. Ironically, these critiques have been met with more testing, not less, and performance on some assessments has been tied to higher stakes such as promoting and retaining both students and teachers (Ball & O'Connor, 2016; Sutter et al., 2020). With so much scrutiny and consequences associated with state achievement testing, it is vital to ensure that students are on track for success well before their testing grades.

Multi-Tiered Systems of Support and Response to Intervention

Early identification of students at risk for academic difficulties was a primary reason for including a response to intervention (RTI) framework in the reauthorization of the Individuals with Disabilities Education Act in 2004. Within an RTI framework, intervention is provided to accelerate the academic progress of students who are at risk academically (Compton et al., 2012). Since then, ESSA (2015) introduced a multi-tiered systems of supports (MTSS) which includes RTI for core academic subjects along with early identification of students who may require additional behavioral and social-emotional learning supports (Al Otaiba et al., 2019; Whittaker & Batsche, 2019).

The use of an RTI model is motivated by educators' desire to ensure that students are on-track or developing at an adequate rate toward the goal of standards proficiency and to pass state tests through early detection of difficulty and immediate intervention (Silberglitt & Hintze, 2005). This early detection and intervention can prevent reading

difficulties for many students (Al Otaiba et al., 2011; Fletcher et al., 2019). Effective RTI models also can differentiate between students who are low achieving due to underlying correlates such as cognitive or neurobiological factors (Fletcher et al., 2005).

Most RTI models consist of similar components. Accurate identification of students at risk for academic difficulty through a universal screener is the first step in the RTI process (Compton et al., 2012). The next step is to provide timely and appropriate interventions based on the data (Whittaker & Batsche, 2019). Then, student progress is monitored frequently, and the data is used to decide the appropriateness and intensity of the interventions (Compton et al., 2012). These steps take place within a multi-tiered framework alongside high-quality classroom instruction (Whittaker & Batsche, 2019). Each component is essential for an effective RTI model.

Universal screening, the first step in the RTI process, is a prerequisite for early identification and intervention (Glover & Albers, 2007). Universal screeners are measures given to all students to identify students at risk for academic or behavioral difficulties and whose progress needs to be monitored (Glover & Albers, 2007; Whittaker & Batsche, 2019). Effective universal screeners are brief, cost-effective, aligned with the constructs or areas of interest, and allow for multiple administrations across a school year, which increases the reliability of estimated change and growth over time (Fletcher et al., 2005; Klingbeil et al., 2015). Curriculum-based measures (CBM) are a popular type of universal screener that have been around for a while. These measures are standardized measures for assessing students' skills in reading, math, spelling, and written expression (Shinn & McMaster, 2019). Measures like curriculum-based measures

allow educators to track student progress and determine if a student is on track for meeting grade level benchmarks (Silberglitt & Hintze, 2005).

The tiered intervention system in RTI is based on increasing the intensity of intervention as students demonstrate increasing need based on universal screening and progress monitoring data (Van Norman et al., 2017). The exact number of tiers in a particular RTI model may vary, but typically begin with core classroom instruction as Tier 1 for all students (Fuchs et al., 2010). Subsequent tiers provide more targeted interventions in smaller groups, with the final tier being special education (Fuchs et al., 2010).

Once a student has been identified and placed into an appropriate intervention, then the student's progress is regularly monitored. Progress monitoring is used for any student receiving an intervention to determine the appropriateness of the current placement and to make decisions about movement through the tiers (Lam & McMaster, 2014). Students are progress monitored after the universal screening has been administered (Fletcher et al., 2005). Linking multiple assessments to the interventions supports the identification of student's responsiveness or non-responsiveness to the attempts to intervene, thus allowing informed decisions regarding the next steps for students (Fletcher et al., 2005).

Universal Screening

As the first step in the RTI process, universal screening is critical for making informed decisions regarding placement and intervention for students. Universal screening involves administering quick, valid measures of academic indicators to all students within a grade, class, school, or district (Hosp & Ardoin, 2008). The measures

are administered with standardized protocols to reduce the impact of test administration on student performance (Hosp & Ardoin, 2008). Since all students are screened, data can be used to place students in appropriate tiers in an RTI model (Whittaker & Batsche, 2019).

Purposes of Screening

Using universal screeners at regular intervals throughout a school year can provide valuable information for detecting, preventing, and intervening for reading or academic difficulties. A universal screening process requires multiple administrations throughout the year, typically in the fall, winter, and spring (Hosp & Ardoin, 2008). As Francis and colleagues (2005) note, no single assessment given at a single point in time is sufficient for making decisions that can have a long-term impact on students. Multiple measure administrations allow educators to review changes in growth and achievement over time, which increases the accuracy of decisions being made such as those informing the identification of reading disabilities (Al Otaiba et al., 2011; Francis et al., 2005; Odegard et al., 2020). Using a valid universal screener multiple times throughout the year provides a more thorough picture of students' performance.

Universal screening is also a means of preventing reading and academic difficulties because of its proactive approach to assessing academic indicators of reading proficiency. Traditional means of identifying student risk have been criticized for their reactionary "wait to fail" approaches in which students were not assessed or provided interventions until after they were struggling (King et al., 2016). Universal screening offers the opportunity to identify students at risk for future reading or academic difficulty

and provide interventions immediately to set them on a typical reading development trajectory, thus preventing the future need for special education (Compton et al., 2012).

Universal screening results are often used to identify students' instructional needs and place them in the appropriate tier or group to accelerate the performance of all students (Hosp et al., 2011; Whittaker & Batsche, 2019). For this reason, it is essential to have a universal screener that accurately measures progress toward academic standards and identifies students who are not on track to meet those standards so they can be matched to the appropriate intervention without the need to wait for them to fall behind (Al Otaiba et al., 2014; Fletcher et al., 2005). Universal screening data inform decisions regarding which interventions to use, the intensity of intervention necessary, and how to use staff to support the implementation of the interventions (Lam & McMaster, 2014). Students who demonstrate the need for the most intensive interventions can move directly to the highest tier of intervention specified in the RTI model (Compton et al., 2012). Most students needing intensive intervention will need it over a sustained period of time (Al Otaiba et al., 2014). It is important to note that universal screening data and RTI models are not just a means of determining eligibility for special education or only for students who struggle (Whittaker & Batsche, 2019). The data should be used to make decisions and differentiate instruction and intervention for all students (Fletcher et al., 2019).

Types of Screening Measures

Universal screeners are designed to target either specific academic skills or established academic standards. Skill-based universal screeners such as Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2020) are quick measures designed to minimize the impact on student instructional time (Ball &

O'Connor, 2016; Van Norman et al., 2017). Standards-based universal screeners typically take longer to administer but assess students' achievement and growth towards grade-level standards (Thomas & January 2021). Although skills-based screeners are more widely used, many states and districts allow standards-based screeners to be included as a first step in the RTI process (Whittaker & Batsche, 2019).

Curriculum-based Measures. Curriculum-based measures (CBM) are short, nationally normed general outcome measures typically administered in a one-on-one setting and designed to assess student performance and progress toward basic skills (Ball & O'Connor, 2016; Shapiro et al., 2006). General outcome measures are single, general tasks that provide an indication of change in a broader outcome such as reading proficiency (Deno, n.d.). Curriculum-based measures, or probes, are administered using a standardized procedure at regular intervals throughout the school year (Kettler & Albers, 2013). There is much evidence to support decision making based on curriculum-based measures (Kettler & Albers, 2013; Reschly et al., 2009).

In 2015, Catts and his colleagues studied the predictive validity of curriculum-based measures in kindergarten to reading outcomes at the end of first grade. They found that all screening and progress monitoring measures were significantly correlated with first grade reading outcome measures. The Letter Name Fluency measure was the strongest predictor of reading outcomes (Catts et al., 2015). They also found that rapid automatic naming and nonword repetition provided unique variance to the outcome measure (Catts et al., 2015).

January & Klingbeil (2020) evaluated the validity and diagnostic accuracy of various reading curriculum-based measures in kindergarten to grade 2. They found that

reading curriculum-based measures have medium to large associations with other similar and more complex measures. For example, Letter Name Fluency and Letter Sound Fluency had large correlations of .52 to .64 with broad reading measures. The measures of phonological awareness, Onset Sound Fluency and Phoneme Segmentation Fluency had slightly smaller correlations, .34 to .51, with the reading outcome measures. They also found that Nonsense Word Fluency and Word Identification Fluency had large correlations with broad reading measures. First grade WIF was also linked to distal reading outcomes in grade 2 in a previous study by Compton and colleagues (2012). January and Klingbeil (2020) concluded that curriculum-based measures can be interpreted as indicators of early reading achievement. However, they noted that more research is needed on the classification accuracy for identifying students at risk.

The most researched curriculum-based measure is the measure of oral reading fluency (ORF). For ORF administration, students are given a grade-level passage and asked to read it aloud for one minute. As the student reads, the assessor marks any errors the student makes. A student's final score is the number of words read correctly per minute and can be compared to national fluency norms (Hasbrouck & Tindal, 2017). A student's accuracy can also be determined. This general outcome measure is a quick check of a student's reading fluency and accuracy at a given point in time (Deno, 2003). ORF measures can be used for benchmarking and are sensitive enough for more frequent progress monitoring (Deno, 2003). ORF measures are sensitive enough to differentiate between typical and at risk readers (Petscher et al., 2019b).

Oral Reading Fluency has also been found to be strongly correlated to state achievement tests. Shin and McMaster (2019) conducted a meta-analysis of the

relationship between ORF and MAZE, a reading comprehension curriculum-based measures, and state achievement tests. They included 61 studies that met their inclusion criteria in their analysis. In grades 1-6, correlations between ORF and state achievement tests ranged from 0.60 to 0.80. They also noted that the correlations decrease as a function of increasing grade level, which indicates that screening materials may need to shift as students move up in grades. Doing so would ensure that the assessments are most appropriate for the student's needs and can promptly support educators in providing the most appropriate interventions. They concluded that CBM reading tasks (ORF and MAZE) are valid indicators of reading comprehension on state tests (Shin & McMaster, 2019). These findings, they noted, were moderated by the type of test (i.e., commercial or state developed) and the time interval between the administrations of the test.

When using curriculum-based measures as part of the universal screening process, research favors the use of multiple indicators over any single measure (Kettler & Albers, 2013). CBM are intended to be quick indicators of risk and, therefore, may lack the diagnostic information necessary for making informed decisions (Ochs et al., 2020). Kettler and Albers (2013) found that including teacher ratings of students' reading ability improved accurate risk classification of curriculum-based measures. In addition, Lam & McMaster's (2014) review supported using multiple measures of word identification, alphabetic principle, fluency, and phonemic awareness in screening batteries. They found that student demographics and vocabulary alone were generally not predictive and that the use of fluency measures, in addition to accuracy-based measures, improved the identification of risk (Lam & McMaster, 2014).

Multiple curriculum-based measures are also predictive of performance on state achievement tests and can be used to predict future performance as early as in grade 1. Shapiro et al. (2006) found that CBM-Reading (CBM-R) or oral reading fluency measures were a strong predictor of success on state achievement tests which is consistent with other studies (Keller-Margulis et al., 2008; Yeo, 2010). Silberglitt and Hintze (2005) found that CBM-R was a strong predictor of success on the state achievement test with a high predictive and concurrent validity. They further suggested that schools set local cut scores on CBM-R that are more predictive of passing the state achievement test to allow schools to determine students who are at risk of not being proficient on the state test as early as first grade (Silberglitt & Hintze, 2005).

Curriculum-based measures used as universal screeners provide a quick estimate of basic academic reading skills. Their brevity and versatility have made them commonly used for universal screening within an RTI framework. Evidence has established their ability to predict distal outcomes, including state achievement tests, which has added to their popularity. However, they are often given in a one-on-one setting, which requires a lot of time and personnel to complete in a timely manner. Also, their brevity does not allow the measures to provide necessary diagnostic information.

Computer-Adaptive Tests. An alternative type of universal screener is the computer-adaptive test. Computer-adaptive tests are typically group-administered, computer scored measures that can estimate students' current instructional needs across a wider range of domains (Ball & Connor, 2016). According to Shapiro and Gebhardt (2012), a computer-adaptive test can identify specific skill areas of strength and weakness by refining test items presented based on the students' responses for more domains than a

traditional CBM. This adaptation is based on the premise that a domain comprises various underlying skills that follow a developmental progression (Shapiro & Gebhardt, 2012). Using item-response theory (IRT), each student receives a unique test (Ochs et al., 2020). Therefore, a computer-adaptive test can provide more diagnostic information than a CBM assessment designed to measure a single skill or a general skill (Shapiro & Gebhardt, 2012). Although computer-adaptive tests have been used as part of assessment programs in schools, there is limited research on their use as universal screeners (Ball & O'Connor, 2016; Sutter et al., 2020).

Ball and O'Connor (2016) examined the utility of a computer-adaptive measure and a CBM for predicting performance on state-level achievement tests and for their ability to identify student risk for academic difficulty. Their measures were the standards-based Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) and the skills-based DIBELS oral reading fluency. Each measure was administered in the spring of grade 2, and both were found to be significant predictors of the fall of grade 3 state achievement test performance. They also found that the use of locally derived cut-scores for the universal screeners improved the classification accuracy of each measure. In their conclusion, they stated that grade 3 state test performance could be accurately predicted by screening students in the spring of grade 2 using MAP and DIBELS oral reading fluency together (Ball & O'Connor, 2016). In 2020, Ochs and colleagues (2020) examined the predictive validity of another computer-adaptive test, the Star Reading assessment (STAR-R) from Renaissance Learning, for predicting performance on a state assessment for students in grades 3 to 5. They found significant

correlations between STAR-R and state tests within the same school year and one to two years later (Ochs et al., 2020).

More recently, Thomas and January (2021) studied the relationship between the MAP and another state assessment. They argued that the urgency required for identifying students at risk for reading difficulties and the importance of accurate classification requires a universal screener that conserves time and resources, a characteristic of computer-adaptive tests (Thomas & January 2021). In their study, they found that MAP scores in the spring of second grade were a statistically significant predictor of the state assessment in the fall of third grade for two cohorts of students using local cut scores (Thomas & January 2021).

Emerging evidence has shown that computer-adaptive tests hold promise for their use as universal screeners and warrant further research. Computer-adaptive tests may improve accuracy in identifying risk and providing diagnostic information that can lead to more targeted interventions (Klingbeil et al., 2015; Ochs et al., 2020). Their ability to adapt items presented to each student and their demonstrated correlations to state assessments further support their potential for use as universal screeners (Shapiro & Gebhardt, 2012; Sutter et al., 2020). Other researchers have noted additional benefits of computer-adaptive tests, such as their cost effectiveness since no printing is necessary for the computer test and automated scoring that eliminates the time needed for grading (Sutter et al., 2020; Thomas & January 2021). These benefits are beyond what traditional curriculum-based measures can provide (Ochs et al., 2020).

Domains of Universal Screeners

Universal screeners for reading are designed to assess student performance across a set of literacy domains and are often based on developmental reading research or state literacy standards. Broad reading constructs of word reading and comprehension broken down into various domains such as phonological awareness, word reading, vocabulary, fluency and comprehension are often included.

Phonological awareness is the ability to hear and manipulate the parts of language from the sentence level to the phoneme level (Wanzek et al., 2020). A subskill of phonological awareness, phonemic awareness is the ability to hear and manipulate individual sounds in spoken words and is a strong predictor of early reading and later reading skills (Peng et al., 2019). Students begin with identifying rhyming words and counting the syllables in words. Then, students proceed to blending onsets and rimes, producing rhyming words, and isolating initial sounds in words for matching sounds. Students then develop the ability to blend, segment, and substitute simple words with up to four phonemes (Wanzek et al., 2020). At the highest or most complex level, students are able to delete phonemes from words, including words with blends. These skills develop in a general progression and overlap rather than in discrete, isolated steps. An inability to segment words or syllables is a strong predictor of early reading difficulty (Ehri, 2004). These emergent literacy skills are most often included in grade-level standards and assessments in kindergarten and first grade but may be used for identifying skill deficits in older students (Catts et al., 2015).

Word reading skills are foundational to reading comprehension and are a critical predictor of reading success (Clemens et al., 2019; Perfetti & Stafura, 2014; Vaughn et

al., 2020). Letter knowledge and an understanding of the alphabetic principle are the first aspects of word reading introduced to young readers and have been shown to be significant predictors of both early and later reading skills (Ehri, 2004; Peng et al., 2019). The alphabetic principle is the knowledge that individual phonemes, or speech sounds, are represented by graphemes or letters (Ehri, 2004). Decoding is the ability to pronounce written words. Students typically follow a developmental progression of word recognition skills that moves from the emergent literacy skills of letter knowledge and letter-sound correspondences to decoding simple words to decoding larger word parts to finally being able to accurately and fluently read (Ehri, 2004). Word reading assessment tasks include letter naming, letter sounds, word reading fluency, and nonsense word fluency.

Vocabulary is another domain included in many universal screeners because of its strength as a predictor of reading comprehension. To be proficient readers, accurate and fluent reading is insufficient. Students must know the meanings of the words being read. Vocabulary both directly and indirectly impacts reading comprehension (Elleman & Oslund, 2019). As a necessary component of reading proficiency at all grade levels, vocabulary has been shown to be a particularly good indicator of state reading test proficiency when compared to other curriculum-based measures assessments (Nese et al., 2008). As students get older and progress through school, vocabulary offers a unique contribution to reading comprehension (Clemens et al., 2019; Peng et al., 2019).

Reading comprehension is a more complex task, and therefore, it is more difficult to assess (Catts, 2018). However, as the ultimate goal of reading, comprehension is a key domain assessed on universal screeners. Reading comprehension is making sense of text

through the coordination of a variety of cognitive processes (Elleman & Oslund, 2019; Silva & Cain, 2015; Wanzek et al., 2020). It involves building a coherent representation of the text by making connections between the ideas and events in the text with prior knowledge and is supported by inference-making. Various strategies may be employed to make those connections including strategies to correct understanding when comprehension breaks down (Catts, 2018; Elleman & Oslund, 2019; Wanzek et al., 2020). Because of the complex nature of reading comprehension, it is difficult to assess in a single measure (Elleman & Oslund, 2019). In fact, research has shown that reading comprehension assessments are not highly correlated with one another (Elleman & Oslund, 2019). Yet despite the challenges, a measure of reading comprehension added to ORF improved the predictability for a state reading test (Shapiro et al., 2008).

Spelling may also be included in universal screeners because of its ability to offer insight into a students' reading development by providing information about their knowledge and ability to apply foundational reading skills (Clemens et al., 2014). Learning to read is intertwined with learning to spell due to the common foundational skills on which both are based (Clemens et al., 2014; Graham & Santangelo, 2014). Therefore, assessing spelling skills can provide information about other reading related skills (Clemens et al., 2014).

The domains assessed or required for a composite reading score will vary across universal screeners and can depend on grade level, timepoint, or purpose. Therefore, special consideration should be given to the domains assessed by a universal screener in addition to the population and setting factors when selecting a universal screener.

Universal Screener Selection

When choosing a universal screener, it is essential to consider the measure's appropriateness for the setting in which it will be used. With the high stakes of identifying students at risk for academic difficulties and deciding which interventions to provide, decisions regarding selecting the measure to use are critical. First, a universal screener should be compatible with the needs of the local school or district (Glover & Albers, 2007). The purpose of the screener, the frequency of administration, and the alignment of domains or constructs being assessed are all essential considerations (Glover & Albers, 2007). The theoretical and empirical support for the measure should also be reviewed to ensure a reliable and accurate measure is chosen (Hosp & Ardoin, 2008). The measure should be contextually and developmentally appropriate for the intended population being screened (Glover & Albers, 2007). This can be decided by reviewing test items if available. Finally, the practical characteristics such as the testing time, personnel required for administration, and the impact on instructional time should also factor into selecting a universal screener (Klingbeil et al., 2015).

Reliability and Validity

In the past few decades, independent research on universal screening measures has progressed to the point that measures with sufficient reliability and validity are available that allow educators to feel more confident about the results and decisions being made based on them (Glover & Albers, 2007). However, most of the research has focused on curriculum-based measures, which have been available for longer. As a newer type of universal screener, computer-adaptive assessments must provide accurate and reliable information to make critical decisions for students, especially those who struggle

or are at risk of struggling (Fletcher et al., 2005). The reliability of a measure refers to the consistency of the results (McKenna & Stahl, 2009). Reliability coefficients are computed and expressed as values ranging from 0 to 1, with coefficients closer to 1.0 being considered more reliable. A measure with a reliability coefficient of .90 is considered to have high reliability.

The validity of an assessment is measured in a variety of ways based on the evidence collected (McKenna & Stahl, 2009). For example, content validity reflects the alignment of the curriculum being taught to what is being assessed. Construct validity is the extent to which the test measures what it is supposed to measure and is often verified by comparing the test to other tests that measure those same constructs. Concurrent validity is determined by comparing the results of the test with the results of an established test given at the same time. The higher the correlation between the scores, then the higher the concurrent validity.

For universal screening purposes, predictive validity is important because the results of the screening measures are often being used to predict future risk. Predictive validity is the ability of one test to predict the performance on a future test (McKenna & Stahl, 2009). Relationships between screeners and other academic outcomes are often determined through correlation or regression analysis (Ball & O'Connor, 2016). Correlations demonstrate the strength of association between the two measures (Ball & O'Connor, 2016). Regression analyses can show the amount of variance in the outcome measure accounted for by the screener. Although these analyses provide evidence of the technical adequacy of results, they are insufficient for providing information regarding

the diagnostic accuracy of the testing results (Kane, 2013). Therefore, an additional set of analyses are necessary.

Diagnostic Accuracy

Diagnostic accuracy provides a statistical means of determining the usefulness of an assessment for screening. The diagnostic accuracy, or classification accuracy, of universal screeners is most often evaluated using methods derived from signal detection theory and is considered to be a form of predictive validity (Petscher et al., 2019a; Trevathan, 2017). Signal detection theory is a method for measuring a system's ability to detect or recognize a signal or event despite background noise (Swets, 1996). In detection, the goal is to detect a signal from the noise. In a recognition task, the goal is to recognize and differentiate two signals (e.g., category A and category B). Signal detection theory allows for the separate evaluation of discrimination (i.e., the degree to which evidence points to the existence of a signal from the noise or signal A from signal B) and the decision process (i.e., the process to determine how strong the evidence must be to choose one signal or the other; Swets, 1996). In education, signal detection theory can be applied to maximize the diagnostic accuracy by accurately identifying students who are at risk or not at risk of future reading struggles on a gold standard measure of reading (Kent et al., 2019).

In education, diagnostic accuracy is a method for establishing the predictive validity of screeners with dichotomous or continuous predictors and dichotomous outcomes that goes beyond establishing a point correlation between a predictor and a criterion (e.g., pass/fail; Ball & O'Connor, 2016; Hosp et al., 2011; Thomas & January 2021). Diagnostic accuracy metrics include population-based sensitivity and specificity

and sample-based positive and negative predictive power (Trevethan, 2017). Sensitivity, the true-positive proportion, is defined as the number of cases that test positive on the screener and positive on the gold standard outcome divided by the total number of positive cases on the gold standard measure. Trevethan (2017) suggests that a more precise definition of sensitivity includes the probability of correctly identifying all those who have a condition or are positive while minimizing the false negative rate. Specificity is defined as the number of cases that test negative on the screener and are also negative on the gold standard outcome divided by the total number of negative cases on the gold standard measure (Trevathan, 2017; VanDerHeyden, 2010). High sensitivity and specificity levels are necessary and should be prioritized for universal screening measures (Kent et al., 2019; Petscher et al., 2019a; VanDerHeyden, 2010). Yet, there is no consensus on the optimal level of sensitivity or specificity in the literature (Kent et al., 2019). Acceptable levels of sensitivity and specificity often range from .70, considered to be an adequate level, to .95, a high level (Catts et al., 2015; Compton et al., 2006; Kilgus et al., 2014; Hosp et al., 2011). Regardless of the level set, a balance between sensitivity and specificity is needed to ensure accurate results (Ball & O'Connor, 2016).

The sample-based metrics of positive and negative predictive power should also be considered when determining the diagnostic accuracy (i.e., classification accuracy) of a screener. These statistics are based on the sample being screened and are influenced by the prevalence of the condition within the sample and base rate effects, the percentage of at-risk students according to the outcome measure (Ball & O'Connor, 2016; Kent et al., 2019). Positive predictive power is a screening test's probability or odds that a positive test finding is truly positive. Positive predictive power is calculated by dividing the

number of individuals indicated as at risk on the screener and the gold standard measure by the number of positive cases on the screening measure (Trevathan, 2017; VanDerHeyden, 2010). In contrast, the negative predictive power is the probability or odds that a negative test finding is truly negative (Trevathan, 2017; VanDerHeyden, 2010). Trevathan (2017) argues that positive and negative predictive power should be used to make decisions at the individual level rather than sensitivity and specificity. He suggests that high sensitivity may not be helpful for definitively saying a condition is present or not. Instead, high sensitivity allows a person to be ruled out as having a condition if the person tests negative on the screener. In contrast, high specificity allows a person to be confident that the condition is present if the test is positive. Moreover, he suggests that high positive predictive power and negative predictive power are more desirable for making decisions at the individual level to minimize false positives, which can be associated with overtreatment and unnecessary costs of interventions (Trevathan, 2017).

Estimating a receiver (or relative) operating curve (ROC) is one of the primary methods for visually evaluating the diagnostic accuracy of a universal screener by graphically separating the discrimination from the decision process (Hosp et al., 2011; Swets, 1996). An ROC curve plots the proportion of true positives against the proportion of false positives across all bias cut points for decision making. The primary objective of ROC is to identify the decision criterion (i.e., cut score on a screener) that maximizes the best balance of true-positive proportion and false-positive proportion, which vary with one another. When the true-positive rate is equal to the false-positive rate, the discrimination is zero, meaning there is no discrimination. When the true-positive rate is

equal to one for all values of the false-positive rate, then the system has perfect discrimination (Swets, 1996).

The area under the curve (AUC) is the most common value for the overall accuracy of a screening measure. It provides the probability of a predictor correctly identifying two random students in a sample (Hosp et al., 2011). The AUC values range from 0.5, equivalent to chance, to 1.0 for a perfect screener. AUC values above .90 represent excellent diagnostic accuracy, while values between .80 and .90 are considered good (Compton et al., 2006). The AUC is also used as an effect size statistic (Hosp et al., 2011). However, it is essential to note that the AUC does not account for the condition's prevalence or the costs of misdiagnosis (VanDerHeyden, 2010).

When making decisions based on universal screeners, considerations should be given to the diagnostic accuracy of the measure, the scope of assessment, and its reliability and validity (Petscher et al., 2019a). In addition, determining the diagnostic accuracy is an important first step in making student placement decisions because it has important implications for allocating resources in a school (Kent et al., 2019; Thomas & January 2021). Therefore, diagnostic accuracy should be determined for universal screeners to maximize the decision making rules and procedures associated with the measures (VanDerHeyden, 2010).

Assessing Diagnostic Accuracy

Researchers have explored the diagnostic accuracy of universal screeners using statistical methods to determine how well do screeners predict outcomes on state assessments. These methods allow researchers to explore beyond the strength of the relationship between two measures based on correlations. In 2005, Silberglitt and Hintze

compared the utility and accuracy of four statistical methods for establishing CBM-Reading cut scores to predict passing or failing the state achievement test. They suggested that establishing local cut scores was a useful alternative to the traditional normative data. Using discriminant analysis, the equipercentile method, logistic regression, and receiver operating characteristic (ROC) curves, they found that a combination of logistic regression and ROC curve analysis were the strongest methods for establishing cut scores that result in the highest levels of diagnostic accuracy. These methods have been used throughout the medical and educational research exploring the diagnostic accuracy of screeners.

Shapiro and his colleagues (2006) applied ROC curve analysis to their work to determine the diagnostic accuracy of curriculum-based measures. After determining a moderate to strong relationship between the two measures across the fall, winter, and spring benchmarks through correlational analysis, they created ROC curves to determine the sensitivity, specificity, positive predictive power, and negative predictive power. They found that reading curriculum-based measures had levels of sensitivity and specificity above .7 based on the local cut scores, which are considered to be acceptable levels for a screening measure.

Following the publication of Silberglitt and Hintze (2005) and Shapiro et al. (2006), other researchers began to examine ways to improve the diagnostic accuracy of universal screeners by the inclusion of additional measures. Adding a screening measure of reading comprehension to ORF resulted in a greater prediction of outcomes on a state reading test (Shapiro et al., 2008). Using logistic regression and ROC curve, the authors examined the diagnostic accuracy of DIBELS ORF and 4Sight Benchmark Assessment

for classifying students as at risk of not passing the Pennsylvania System of School Assessment (PSSA) for 1000 students in grades 3 through 5. Speece and colleagues (2010) sought to identify a screening battery that would accurately identify students in need of more intensive intervention. Using a variety of measures designed to assess reading comprehension, listening comprehension, word recognition, decoding, phonological processing, and spelling, they assessed 230 fourth-grade students from 15 parochial schools. An additional teacher rating of students' abilities was included. Using regression analyses to determine a subset of predictors followed by logistic regression and ROC curves, they determined that risk was best identified through measures of reading comprehension, word reading fluency, and teacher ratings. When combined, the three measures resulted in an area under the curve of .90. They concluded that a multivariate approach to universal screening is necessary for accurate identification, a finding supported by other research (Johnson et al., 2010; Keller-Margulis et al., 2008; Kent et al., 2019; Kettler & Albers, 2013).

In 2012, one of the first direct comparison of curriculum-based measures and computer-adaptive assessments was published by Shapiro and Gebhardt. Again, the researchers began with establishing predictive validity by showing a moderate to strong relationship between the two types of assessments and a state assessment across three assessment timepoints as demonstrated by correlations. To compare the diagnostic accuracy of the two types of measures in their within-subjects design, they used cross-tabulations to identify the cut scores associated with the 16th percentile, a cut point associated with marking proficient/not proficient on the state test. Then, diagnostic

accuracy metrics were calculated and compared against the standard level of .90 as suggested by Johnson et al., (2009).

Setting universal screener cut scores for comparison across different measures on different scales has also been approached in various ways. Like Shapiro and Gebhardt (2012), Clemens et al. (2015) also used cross-tabulations and a predetermined 40th percentile to set cut scores before computing the diagnostic metrics when comparing CBM, Star Early Literacy, and norm-referenced tests. Ochs and colleagues (2020) chose to use the cut score associated with predetermined percentile ranks of 25th and 40th when examining the diagnostic accuracy of Star Reading for predicting outcomes in the state reading test. Others have chosen to set the sensitivity to a desirable level of .90 when using different assessments, and then use the resulting cut scores pulled from the ROC curves to determine diagnostic accuracy (Kent et al., 2019; Klingbeil et al., 2015; Klingbeil et al., 2018).

Another statistical method that can support the goal of understanding and quantifying the diagnostic accuracy of multiple assessments while also determining the importance of the predictor is the comparison of the areas under the curve (AUC) derived from ROC curves (Kuhn & Johnson, 2016). In 1982, Hailey and Long posited that the AUC, as it represents the probability of correctly identifying a randomly selected case, is similar to a Wilcoxon statistic, a test used to determine how useful a discriminator is. Using this relationship, the AUC value and its standard error (*SE*) can be computed from observed data and then compared.

With the available statistical methods for building predictive models commonly used in research, it is possible to quantify the diagnostic accuracy of the predictions being

made using universal screener. Diagnostic accuracy allows researchers to move beyond the question of whether or not a measure is a strong predictor of another and into a discussion centered on how accurately a measure is able to classify students at risk. It further allows for the quantification of how well a model will perform with future data to make an educational decision of risk to triage students for intensified instruction and intervention (Kuhn & Johnson, 2016). However, few studies have been undertaken to evaluate and compare the diagnostic accuracy of commonly used computer adaptive tests used to screen students for risk across the country.

Research Problem

This study examined the predictive validity and diagnostic accuracy of three computer-adaptive universal screeners. Predictive validity is one indicator of technical adequacy established by correlating one measure to another (Ball & O'Connor, 2016). However, predictive validity alone is insufficient to inform the selection of tests for universal screening (Klingbeil et al., 2018). Diagnostic accuracy provides information beyond the strength of the relationship between two measures by providing evidence that supports the accuracy of the decisions being made (Klingbeil et al., 2018; Thomas & January, 2021). Much of the initial research on universal screeners focused on improving the measures themselves rather than the utility and predictability of the measures for other meaningful academic outcomes (i.e., state achievement tests; Hosp et al., 2011; Ochs et al., 2020). Also, a majority of the current supporting evidence has shorter time intervals between the administration of each measure, with most studies using three to six-month intervals between universal screening and achievement test administration

(Hosp et al., 2011). Predictive validity and diagnostic accuracy of each computer-adaptive universal screener was examined in this study.

By using three different computer-adaptive universal screeners, this study adds to the growing body of research on the utility and effectiveness of computer-adaptive measures for universal screening by comparing universal screening data from the spring of grade 2 to state achievement test performance in the spring of grade 3 for students who took one of the three universal screeners. If a universal screener can predict performance before a state achievement testing year, it will increase intervention opportunities (Ochs et al., 2020). Earlier intervention could lead to better outcomes.

Specifically, this study compared the ability of universal screening data obtained from three computer-adaptive universal screeners administered at the end of the second grade to predict a child's likelihood of passing a state reading achievement test given at the end of third grade and the accuracy of those decisions. Grade 2 is the last opportunity to intervene before the mandated state testing begins in grade 3 (Ochs et al., 2020). By grade 2, students have already had up to two years of formal reading instruction. Cross-grade prediction can open additional opportunities for intervention and support in-between the grade levels or allow intensive intervention to begin immediately at the start of the next school year (Ball & O'Connor, 2016). Proficiency on state tests is one outcome of particular importance to schools. It has the potential for identifying students at risk for academic failure before the testing year, which can maximize intervention time (Keller-Margulis et al., 2008).

This study addressed the current gap in research on cross-grade level prediction of performance on state achievement tests using computer-adaptive universal screening

measures. Including three computer-adaptive universal screeners that differ in domains assessed, administration times, formats, and scores reported in this study provides a unique opportunity to examine their predictive validity and diagnostic accuracy for a single outcome measure not specific to one state, an analysis that is lacking in the current literature. Of the three computer-adaptive universal screeners in this study, only the Istation Indicators of Progress assessment includes a measure of reading fluency, a measure that is a strong predictor of reading outcomes. MAP and Star Reading do not. All three universal screeners included in this study have vocabulary and reading comprehension measures, but only one includes a spelling measure (i.e., Istation). These different reading domains assessed by default could impact the diagnostic accuracy of the universal screeners for a state standardized test.

The current study also adds to the literature by examining three different computer-adaptive universal screeners in relation to one widely available state achievement test. Many studies include only one of the computer-adaptive measures as a predictor for a single, state-specific assessment (e.g., Ball & O'Connor, 2016; Clemens et al., 2015; Kettler & Albers, 2013; Ochs et al., 2020). One study did examine the predictive validity of Istation and Star Reading, but the outcome measure was an achievement test specific to one state (Sutter et al., 2020). The outcome measure in this study is a nationally normed, widely available measure, the ACT Aspire. One Istation internal study that linked Istation to the ACT Aspire identified a high correlation of 0.71 when using the spring of Grade 2 Istation scores, and the middle-of-the-year Grade 3 ACT Aspire scores (Patarapichayatham & Locke, 2020). Star Reading ($r = .84$) and MAP

(r = .80) have also been shown to be correlated to ACT Aspire based on research published by the publishers (NWEA, 2020; Renaissance Learning, 2017).

The current study attempted to expand the research by comparing the predictive validity and diagnostic accuracy of three computer-adaptive measures to a state assessment that was not created for a single state alone.

Specifically, the study will address the following research questions:

1. To what extent is performance on each computer-adaptive universal screener in the spring of 2nd grade predictive of performance on the state-mandated achievement test in the spring of 3rd grade?
2. Which of the three computer-adaptive universal screeners has the best diagnostic accuracy based on their area under the curve?
3. Using a direct route screening approach, what is the diagnostic accuracy of each computer-adaptive universal screener on a state-mandated achievement test using one of the most commonly-used statistical methods for diagnostic accuracy (i.e., ROC curve analysis) and based on (a) cut score associated with the 40th percentile, (b) cut score when sensitivity is set to .90, and (c) cut score that maximizes both sensitivity and specificity?

To address research question 1, the correlations between each computer-adaptive universal screener and the ACT Aspire were examined to establish the strength of association between the measures. This preliminary analysis is consistent with the research literature as a standard practice for establishing each screener as a valid measure for predictive purposes (Clemens et al., 2015; Klingbeil et al., 2018; Ochs et al., 2020; Patarapichayatham & Locke, 2020; Speece et al., 2010; Sutter et al., 2020; Thomas &

January 2021). In addition to correlations, logistic regression models were run for each screener.

To address question 2, receiver operating characteristic (ROC) curves will be modeled for each computer-adaptive universal screener. The procedure is a common procedure for diagnostic accuracy research (Ball & O'Connor, 2016; Ochs et al., 2020; Silberglitt & Hintze, 2005; Thomas & January, 2021; VanMeveren et al., 2020). The universal screener scores will be entered as continuous variables while the outcome variable will be dichotomous (0 = pass, 1 = fail). Modeling the data in this manner allowed for the areas under the curve be calculated. The AUC is a summary statistic for overall diagnostic accuracy across all possible decision criteria (i.e., cut scores; Swets et al., 2000). A comparison of the AUC between each of the three measures will be performed using established procedures (Hanley & McNeil, 1982; Kuhn & Johnson, 2016).

Finally, for research question 3, ROC curves for each computer-adaptive universal screener were used to identify the cut score associated with the 40th percentile on each measure, the cut score that maximizes sensitivity, and the cut score that maximizes both sensitivity and specificity. In this regard, ROC curves have been found to provide the most flexibility and highest levels of diagnostic accuracy for setting standards and establishing cut scores (Silberglitt & Hintze, 2005). For part a, the cut score associated with the 40th percentile will be used. The 40th percentile has been suggested for benchmarking to allow for the identification of more students who potentially may be at risk (Ochs et at., 2020). For part b, the sensitivity rate will be set at .90, as suggested by Compton et al. (2006). For part c, the findings for sensitivity and specificity will be set at

.80 and .70 respectively, based on the suggestions for maximizing sensitivity and specificity from a meta-analysis by Kilgus and colleagues (2014). Once the cut scores for each parameter were identified, then the sensitivity, specificity, positive predictive power, negative predictive power, and overall classification rate was calculated for each cut score and the resulting contingency matrices were created.

CHAPTER III: METHOD

The following methodology was implemented to investigate the predictive validity (i.e., research question 1) and diagnostic accuracy of the computer adaptive tests (i.e., research questions 2 and 3).

Participants

This study uses extant data from Arkansas collected during the 2017-2018 and 2018-2019 school years. The analytic sample includes students enrolled in Grade 2 during the 2017-2018 school year and enrolled in Grade 3 during the 2018-2019 school year in 417 elementary schools in the state. Table 1 shows the demographics of the analytic sample by screener and total analytic sample. All students included in the analytic sample had a universal screener score from Istation, MAP, or Star Reading in second grade and a state ACT Aspire score from third grade. The resulting analytic sample includes 23,292 students.

Table 1

Sociodemographic Characteristics of Students in the Analytic Sample

Characteristic	IStation (n = 4852)		MAP (n = 12,410)		Star Reading (n = 6030)		Total (n = 23,292)	
	n	%	*n*	%	*n*	%	*n*	%
Race								
White	3311	68.2	7116	57.3	3760	62.4	14187	60.9
African American	488	10.1	2192	17.7	1617	26.8	4297	18.4
Hispanic	745	15.4	2222	17.9	444	7.4	3411	14.6
Other groups	308	6.3	880	7.1	209	3.5	1397	6.0

Table 1

Characteristic	IStation (n = 4852)		MAP (n = 12,410)		Star Reading (n = 6030)		Total (n = 23,292)	
	n	%	*n*	%	*n*	%	*n*	%
Gender								
Male	2457	50.6	6216	50.1	3046	50.5	11719	50.3
Female	2395	49.4	6194	49.9	2984	49.5	11573	49.7
Free & Reduced Lunch								
Yes	3159	65.1	7325	59.0	4256	70.6	14740	63.3
No	1693	34.9	5085	41.0	1774	29.4	8552	36.7
SPED								
Yes	693	14.3	1504	12.1	728	12.1	2925	12.6
Limited English Proficiency	547	11.3	1814	14.6	222	3.7	2583	11.1

Note. SPED = Special education services

The 417 schools from which our analytic sample was drawn represent a diverse range of academic performance. Table 2 displays the breakdown of schools by the letter grade assigned to each school through the state rating system. The state's rating criteria used a multiple measure approach based on each school's academic achievement, school value added growth scores which includes English Learner progress, adjusted cohort graduation rates, and indicators of school quality and student success (e.g., on-grade level reading, ACT scores, attendance, proficiency on state tests, GPA, and on-time credits) (Arkansas Department of Education, 2018).

Missing Data

Due to the large sample available for this study, students missing either their spring 2018 universal screening score, their spring 2019 state achievement test score were dropped from the sample. Approximately 32% of the original sample was missing either a universal screening score or the state achievement test score. Results from the analytic sample are reported.

Table 2

Descriptive Statistics of School Letter Grades by Screener

Screener	A		B		C		D & F		NA		Total Schools
	n	%	*n*	%	*n*	%	*n*	%	*n*	%	*n*
Istation	9	10.5	25	29.1	47	54.7	5	5.9	0	0	86
MAP	44	21.7	60	29.6	59	29.1	40	19.7	0	0	203
Star Reading	13	10.2	39	30.5	40	31.3	35	27.4	1	0.7	128

Measures

Istation's Indicators of Progress Reading

The Istation's Indicators of Progress Early Reading is a computer-adaptive assessment of critical reading skills for Pre-Kindergarten (Pre-K) through Grade 3 that provides an overall reading ability score. The test measures student abilities in phonemic awareness, alphabet knowledge, fluency with connected text, vocabulary, and comprehension through various subtests (Mathes, 2009). The default subtests for Grade 2 are vocabulary, reading comprehension, spelling, and text fluency (Mathes et al., 2016).

Item difficulty within each domain range from Pre-K through Grade 5. Items also represent varying difficulty levels, with items corresponding to easy, moderate, and hard for each grade (Mathes et al., 2016).

Istation is administered to whole classrooms during monthly assessment periods no longer than 40 minutes. The assessment is presented in a game-like format, complete with art graphics and an animated announcer, thus allowing students to feel like they are playing a game rather than taking a test. During the first administration of the assessment, item difficulty defaults to the student's grade level. Once the student responds to the first item, the assessment presents subsequent items based on the individual student's performance, resulting in items of appropriate difficulty being presented (Mathes et al., 2016). For later administrations of the assessment, the starting point corresponds with the student's ability level as determined by previous administrations.

Once an assessment is complete, the data is immediately available to the teacher, thus allowing timely interpretation and action. Istation uses a measurement index with equal intervals called ability indices. Each ability index is assigned a numeric value that is identical across all grade levels (Mathes et al., 2016). For example, a first grade student and a fourth grade student who score 220 are considered to be performing at the same level. These indices allow student progress to be tracked across multiple grade levels. In addition to the ability index score, national norms are provided for Pre-K to Grade 3 for comparing students' scores to a nationally representative sample. Students with reading comprehension scores are given a Lexile reader measure for instructional purposes (Mathes et al., 2016). Teachers are also provided links to instructional resources such as instructional plans and materials based on a students' assessment performance.

The Istation is a reliable and valid measure. The test-retest reliability based on the Pearson product-moment correlation coefficients across multiple administrations of the measure throughout the year ranged from 0.927 to 0.970 in an internal study with 416 participants (Mathes, 2009). The marginal reliability, the item response equivalent to classical internal consistency, for any testing administration is set to approximately 0.90 (Mathes et al., 2016). Construct validity was established by testing items in mock-up form and later via computer. Item banks were built initially based on previous work by one of the assessment's authors, Torgesen, to determine how to best measure the constructs within each domain. The resulting pool of items was then calibrated under a 2-parameter IRT model. Items with unacceptable fit statistics were removed from the item bank. Concurrent validity was established by computing the Pearson correlation coefficients between the Istation and other established reading measures such as DIBELS, the Texas Primary Reading Inventory, and the Iowa Test of Basic Skills. Correlation coefficients ranged from 0.48 to 0.83 for DIBELS and 0.82 to 0.90 for the ITBS. To establish the predictive validity for Istation, the Pearson correlations between Istation and the Texas Assessment of Knowledge and Skills (TAKS) were computed. The resulting correlations ranged from 0.695 to 0.741, with the Istation demonstrating stronger correlations to the TAKS than DIBELS, which ranged from 0.450 to 0.630 (Mathes et al., 2016). Istation has also shown strong correlations (0.73) to ACT Aspire and other standardized measures (Cook & Ross, 2020; Mathes et al., 2016; Pataraphichayatham & Locke, 2020).

MAP Reading

The Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) is an untimed, iterative computer-adaptive measure of a student's academic performance (NWEA, 2019). MAP can be administered up to four times per year (i.e., fall, winter, spring, optional summer), and each administration takes approximately one hour to complete. Reading is one of the four content areas assessed with MAP. Each content area is organized into instructional areas and subareas. For MAP for Primary Grades, the instructional areas include comprehension, concepts of print, phonics, phonological awareness, vocabulary and word structure (NWEA, 2011). For the MAP for grades 2 – 5, the instructional areas include Literary Text, Informational Text, and Vocabulary, which address the areas of reading comprehension, understanding of genres, and vocabulary. With an item bank of over 42,000 items aligned to various content standards, NWEA can align its assessment to each state's standards by selecting items from the bank that align with the state's standards and writing additional questions if necessary. This process makes MAP a popular choice for universal screening, as evidenced by its presence in all 50 states (NWEA, 2019).

MAP assessments are administered online but require a proctor to monitor and control the student testing. Students log in to the assessment using a unique testing session name and password. Once they have selected their names, the proctor confirms their tests, and students begin the assessment. The first time a student takes the MAP assessment, the first question defaults to the student's grade level. Once the student responds, the remaining test items are selected based on the student's responses. MAP employs a "rapid guessing" alert that notifies the proctor if a student is moving too

quickly through the test to decrease the likelihood of invalid results due to lack of student engagement. The proctor can work to re-engage the student and then resume the student's test. The MAP for Primary Grades (K-2) Reading test consists of 54 items while the MAP for grade 2 – 5 consists of 40 items (NWEA, 2019).

Students' scores, made available within twenty-four hours, are reported as continuous RIT scores that range from 100 to 350. This equal-interval scale aligns vertically and continuously across grade levels (NWEA, 2019). RIT scores are a unique measurement scale based on the one-parameter Rasch IRT model. The final ability estimate (i.e., RIT score) is computed based on the maximum-likelihood algorithm that places the student's score on the RIT scale. The RIT score is associated with a percentile ranking that shows how the student performed compared to students from the norming group.

The reliability of the MAP assessments was established by test-retest and marginal reliability. Test-retest reliability for MAP was established through a mix of traditional test-retest reliability and alternate forms of reliability due to the adaptive nature of the measures. To estimate the test-retest with alternate forms reliability, the Pearson correlation was computed for consecutive administrations of the test (e.g., Fall 2017 to Winter 2017, Winter 2017 to Spring 2018; NWEA, 2019). The reported coefficient for grade 2 Reading ranged from 0.847 to 0.867 across the school year, indicating that MAP has strong test-retest with alternate forms reliability. Marginal reliability was estimated for MAP to evaluate the measures' internal consistency due to the tests' adaptive nature. The marginal reliability for grade 2 Reading was 0.965, suggesting that MAP has a high level of internal consistency.

The validity of the MAP assessments was established using content validity, concurrent validity, and classification accuracy. To determine the content validity of the measures, the test items were aligned to the Common Core State Standards (CCSS), and it was found that the items were 94.7% aligned to the CCSS. Concurrent validity was established by determining the Pearson correlation coefficient between MAP and state tests administered at roughly the same time. The correlation coefficients ranged from 0.68 to 0.80 for MAP Reading and state tests which fall within the acceptable range as determined by the National Center on Response to Intervention. Finally, classification accuracy indicates whether the MAP assessments are good predictors of student performance on their state tests. According to the MAP Technical Report, MAP showed an overall accuracy rate of 83% for reading, which indicates that MAP is a reliable and valid predictor of student performance on several state tests (NWEA, 2019).

Renaissance Star Reading

The Star Reading assessment is a 34-item standards-based adaptive test aligned with national and state curriculum standards (Renaissance Learning, 2022). The purpose of the assessment is to quickly and accurately measure students' reading skills in five domains (i.e., word knowledge and skills, comprehension strategies and constructing meaning, analyzing literary text, understanding the author's craft, and analyzing an argument and evaluating text) containing ten skill sets (i.e., vocabulary strategies, vocabulary knowledge, reading process skills, constructing meaning, organizational structure, literacy elements, genre characteristics, author's choices, analysis, and evaluation; Renaissance Learning, 2022). Each skill set can be broken down into 36 general and 470 discrete skills (Renaissance Learning, 2022). Using a procedure known

as adaptive branching, students are tested on items that uniquely match their instructional levels.

Star Reading is administered using standard administration procedures to ensure consistency and comparability of results. The frequency of the administration of Star Reading depends upon the purpose for which it is being used. Typically, schools administer the test two to five times a year. Each administration takes an average of fewer than 20 minutes to administer. Although there is no time limit on the total testing time, there is a 60 – 120 second time limit on test items. Each test item is designed to test a specific skill and has undergone a review process to ensure the item meets content specifications. The broad range of test items allows Star Reading to be appropriate for grades K-12.

A student's performance is reported using several estimates and metrics. The Scaled Scores indicate a student's overall performance and can be compared across the year and grade levels. For schools that use the Star Early Literacy assessment, a Unified Scale Score was developed as a common scale to report scores on both tests starting in 2017-2018. Grade Equivalents and Normal Curve Equivalents are also reported. Because the Star Reading test adapts to the student's instructional levels, it can also estimate students' oral reading fluency and instructional reading levels. Students' growth across the year is also reported using a norm-referenced value of student growth called the Student Growth Percentiles.

The reliability of Star Reading was estimated using generic reliability, split-half reliability, and alternate form (i.e., test-retest) reliability. The generic reliability coefficients ranged from 0.94 to 0.96, indicating high reliability. Split-half reliability was

calculated due to the adaptive nature of the test. Scores were compared for the even- and odd-numbered items separately then the correlations between the two sets of scores were corrected to the length of the full 34-item test. The split-half reliability estimates ranged from 0.94 to 0.96 with an overall reliability coefficient of 0.98, indicating high split-level reliability. Alternate reliability estimates were calculated using two different test administrations that avoided repeated test items during the second test. The overall reliability of the scores ranged from 0.81 to 0.87, with an overall reliability of 0.94 (Renaissance Learning, 2022).

Test validity for Star Reading was determined using construct validity, concurrent validity, and predictive validity. Construct validity was estimated by comparing the Star Reading test to the Degrees of Reading Power comprehension test. The comparison resulted in a 0.96 correlation indicating that the two tests measured the same constructs. Confirmatory factor analysis of Star Reading also suggested that the assessment measured a single construct as indicated by the comparative fit index, the goodness of fit, and normed fit index indices values close to 1. Concurrent and predictive validity were investigated by comparing Star Reading to other tests, including state achievement measures, curriculum-based measures, and other norm-referenced tests (see Renaissance Learning, 2022 for a complete list). The average concurrent validity coefficient was 0.74 for external measures administered in the spring, with a range of 0.72 to 0.80 within each grade. The average predictive validity value was 0.71 for external measures administered in fall and spring, ranging from 0.69 to 0.72 in grades 1 – 6. When calculating the concurrent and predictive validity coefficients between Star Reading and state achievement tests, the average concurrent validity coefficient was 0.73, and the average

predictive validity coefficient was 0.68. Overall, Star Reading is reported to have strong reliability and validity as a measure of reading comprehension (Renaissance Learning, 2022).

ACT Aspire

The ACT Aspire Summative Assessment program (ACT, 2020) is the state achievement test given to all students in Arkansas beginning in the third grade. The assessment aims to measure achievement and growth toward college and career readiness standards and includes reporting categories that align with state standards. The reporting categories provide detailed information regarding student performance on key skills. This actionable information can then be used for instructional planning, evaluating program effectiveness, comparing student performance to national norms, and determining a student's progress toward being prepared for high school coursework. The pass rate for 3rd grade in the 2018-2019 school year was 43% (ACT, Inc., 2019).

Student performance is aligned with performance-level descriptors (PLDs), which are descriptions of student performance within and across grade levels. These PLDs are organized into Level 1 - In Need of Support, Level 2 - Close, Level 3 - Ready, and Level 4 - Exceeding. Grade level benchmark cut scores for the Ready category were derived by working backward from the ACT College and Career Readiness scores. The ACT Aspire scores are vertically linked across the grade levels within each subject.

The ACT Aspire test uses a variety of question formats. The assessment contains selected-response items, constructed-response items, and technology-enhanced items when administered on the computer. Selected-response items require the test taker to select one correct response from the choices given. For constructed-response items,

students are instructed to generate a response to the question or prompt by typing into a text box. The technology-enhanced items use various computer features to present questions or scenarios in ways that are difficult to do with the traditional paper format. Student responses may include constructed responses or choosing the correct answer. ACT Aspire replaces technology-enhanced items with selected-response questions for the optional paper testing format. Questions align to Webb's Depth of Knowledge levels to ensure cognitive complexity.

The English Language Arts (ELA) test consists of reading, writing, and English subtests. Scores on the subtests combine to create a composite score for ELA. For the English test, students make revisions and editorial decisions within the context of short passages. The purpose of this subtest is to demonstrate an understanding of writing conventions and strategies while also maintaining style and voice. The reporting categories for this subtest include writing production, writing knowledge, and conventions of standard English. Test items include a variety of texts (e.g., essays, paragraphs, sentences) and different genres. Text length and topic are chosen based on grade-level appropriateness.

The raw score reliability reported for third grade reading was 0.85-0.87 (Patarapichayatham & Locke, 2020). The scale score reliability and standard error of measurement (SEM) were 0.82 – 0.84. ACT Aspire has been reported to correlate with the Partnership for Assessment of Readiness for College and Careers (PARCC) and the ACT test, with a correlation of 0.80 for third grade.

Procedures

This study examined extant data collected from students enrolled in second grade in the spring of 2018 and third grade in the spring of 2019. One of the three computer-adaptive universal screening measures was administered to each second grader in the spring of 2018. The state achievement test was administered in the spring of 2019. The computer-adaptive universal screening measures and the state achievement measure were administered by school staff following district protocols. No fidelity data are available for the measures; however, computer-adaptive tests are designed to reduce variability due to test administration (Shapiro & Gebhardt, 2012).

Data Analysis

For question 1, to establish the strength of the relationship between each computer-adaptive universal screener and the ACT Aspire, Pearson correlations were calculated using SPSS Version 28. Correlations are a common first step in diagnostic accuracy research to establish the predictive relationship between a screener and an outcome measure (Ball & O'Connor, 2016; Kent et al., 2019; McComas et al., 2015; Sutter et al., 2020; Thomas & January, 2021). In addition to correlations, separate logistic regression models were run for each individual predictor to calculate the statistical significance of each screener and the amount of variance accounted for by each screener using the pseudo-R^2 indices. Model fit was also computed. For this analysis, each predictor was entered as a continuous variable and performance on the outcome measure was dichotomized with 0 = no risk and 1 = at risk.

For question 2, to statistically compare the performance of the three computer-adaptive measures, ROC curves were constructed and the AUC for each measure

compared using the approach of Hanley and McNeil (1982). This comparison tests the significance of the difference between the areas that lie under the curve for ROC curves derived from independent samples. This type of statistical analysis is common in medical research, but is rare in diagnostic accuracy research in education (DeLong et al., 1988; Hajian-Tilaki, 2013; Hanley & McNeil, 1983; Zou et al., 2007). In fact, only Catts et al., (2015) included the comparison of the areas under the curve for screening measures in their study of universal screening and progress monitoring for early identification of reading disabilities in kindergarten.

Although overall classification accuracy is essential for universal screeners, universal screening aims to maximize the identification of at-risk students to provide intervention promptly. Therefore, to address question 3, we compared a series of diagnostic accuracy metrics for three cut scores determined by a) the 40^{th} percentile, b) when a sensitivity threshold of .90 was set for each of the measures, and c) when both sensitivity and specificity was maximized using ROC curves. As stated previously, ROC curve analysis is a common practice in diagnostic accuracy research and has been found to produce stronger results than other approaches such as discriminant analysis and equipercentile analysis (Silberglitt & Hintze, 2005). Therefore, ROC curves were used to establish cut scores based on earlier described criteria and calculate the resulting diagnostic accuracy metrics of sensitivity, specificity, positive predictive power, negative predictive power, and overall classification rate.

CHAPTER IV: RESULTS

Preliminary analyses consisted of descriptive statistics, including means, standard deviations, range, skewness, and kurtosis for each universal screener and the outcome measure. Descriptive statistics are presented in Table 3. The values of skewness and kurtosis were within the normal range. Sample size varied across screeners from 4,852 students for Istation to 6,030 for Star Reading, to 12,410 for MAP. Average performance on the ACT Aspire ranged from 417.4 for the sample of participants who completed the Star Reading to 417.5 for the sample of participants who completed the Istation to 418.0 for the sample of participants who completed the MAP. Performance on the ACT Aspire ranged from 405 to 435 across the three samples.

Table 3

Descriptive Statistics for ACT Aspire and Each Universal Screener

Measure	*n*	*M*	*SD*	Minimum	Maximum	Skewness	Kurtosis
ACT Aspire	23292	417.75	5.47	405	435	.107	-.506
IStation	4852	237.70	18.05	151.13	341.89	-.239	.844
MAP	12410	188.09	15.13	118	232	-.463	.216
Star Reading	6030	461.48	131.99	52	920	-.850	.528

Note: The mean (*M*) of each measure is reported on the scale unique to that measure.

Predictive Validity

Research question 1 addressed the predictive validity for each of the universal screening instruments. This was initially achieved by exploring the strength of the relationship between performance on each of the universal screening instruments and

scores on the ACT aspire. To examine the strength of the relationship between each screener and the outcome measure, Pearson's correlations were calculated. Table 4 displays the correlational results for each computer-adaptive universal screener and the statewide achievement test. Each computer-adaptive universal screener was significantly correlated to the statewide achievement test ($p = .01$). Yet, the correlation for Star Reading was considerably lower than the correlation between Star Reading at the beginning of grade 3 and the ACT Aspire ($r = .84$) as reported by Renaissance Learning (2017). Istation's correlation was slightly higher than the correlation between Istation and the ACT Aspire ($r = .71$) as reported by Patarapichayatham and Locke (2020). MAP's correlation was slightly below the correlation between MAP scores at the beginning of grade 3 and the ACT Aspire as reported by the NWEA Technical Report (NWEA, 2020).

Table 4

Correlations between ACT Aspire and Computer-adaptive Universal Screeners

	ACT Aspire
Istation	.781**
MAP	.769**
Star Reading	.537**

Note: **Correlation is significant at the 0.01 level (two-tailed).

To further address research question 1, binary logistic regression analyses were conducted for each universal screener to determine if each universal screener was a significant predictor of performance on the ACT Aspire and to approximate the amount of variance accounted for by each screener. For the logistic regression analyses, performance on the ACT Aspire served as the outcome measure. It was coded as a 0 for

achieving a passing score on the ACT Aspire and as a 1 for receiving a score that was deemed as not passing based on performance criteria published by ACT.

As can be seen in Table 5, results show that all universal screeners were significant predictors of pass/fail on the ACT Aspire. The Istation ($\chi^2 = 2582.47$, p = .000) explained approximately 55.4% of the variance in the ACT Aspire, with the predictor accurately classifying 80% of cases. Nagelkerke is considered to be pseudo-R^2 values. The MAP ($\chi^2 = 6474.40$, p = .000) explained approximately 54.3% of the variance in the ACT Aspire and accurately classified 79% of cases. Finally, Star Reading ($\chi^2 = 1244.24$, p < .001) explained approximately 25.2% of variance and classified 70% of cases accurately. All models were a good fit for the data as indicated by the Chi-square tests.

Table 5

Results of Logistic Regression Analysis for Each Screener

Screener	ß	Standard Error	Wald	*p*	Nagelkerke R^2	CA	SE	SP
Istation	-.142	.004	1128.61	<.001	.554	.80	.82	.78
MAP	-.154	.003	3050.97	.000	.543	.79	.81	.77
Star Reading	-.009	.000	852.30	<.001	.251	.70	.74	.65

Note. CA = classification accuracy; SE = sensitivity; SP = specificity

Diagnostic Accuracy

To answer research question two, we calculated diagnostic accuracy metrics for each computer-adaptive universal screener administered in the spring of Grade 2. First,

ROC curve analyses were conducted for each screener to calculate the area under the curve (AUC). Table 6 shows the AUC value, standard error, and 95% confidence interval for each universal screener.

Table 6

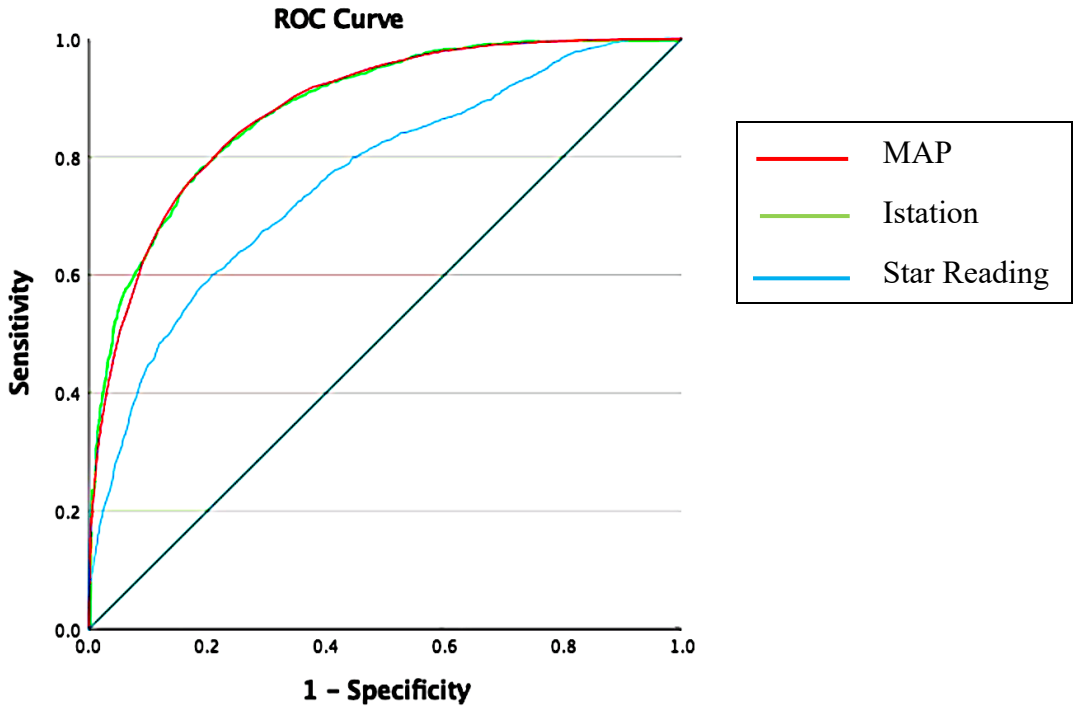
Area Under the Curve by Universal Screener

Measure	Area Under the Curve	Standard Error	95% Confidence Interval Lower Bound	Upper Bound
Istation	.885	.005	.876	.894
MAP	.881	.003	.875	.887
Star Reading	.760	.006	.748	.772

Using the method for comparing AUCs for independent samples outlined by Hanley and O'Neill (1982), each pair of computer-adaptive universal screeners' AUC values was compared for statistically significant differences between the screeners. A critical ratio z is calculated between the two AUC values, and a value greater than 1.65 is considered significant for a one-tailed comparison (Catts et al., 2015). The difference between the AUC for Istation and STAR was statistically significant ($z = 16.31$, $p < .00$). The difference between the AUC values for MAP and Star Reading was also statistically significant ($z = -19.34$, $p < .00$). However, the AUC values for Istation and MAP were not statistically significantly different from one another ($z = 0.80$, $p = .21$). These findings indicate that Istation and MAP were more accurate predictors of passing the ACT Aspire than the Star Reading test, yet do not reliably differ from one another. Figure 1 shows the ROC curves of each universal screener. Appendix A contains Figures A1, A2, and A3 which show the individual ROC curves for each universal screener.

Figure 1

Comparison of the ROC Curves for the Three Universal Screeners



To address the third research question, ROC curve analyses were conducted, followed by a series of crosstabulations to calculate the common diagnostic accuracy statistics of sensitivity, specificity, positive predictive power, and negative predictive power. The results are grouped by the selection process used to establish cut scores for risk on each of the three universal screening instruments. For the first set of analyses, a cut score of risk was set based on performing at the 40th percentile on each of the universal screening instruments. For the second set of analyses, sensitivity was set to .90

and the corresponding cut score was used for each of the universal screeners. In the final set of analyses, a cut score was obtained that maximized sensitivity and specificity for each of the universal screeners. Specificity and sensitivity values ≥ .90 are considered to be excellent, while values ≥ .80 are considered to be good. Values ≥ .70 are considered to be acceptable while values below .70 are considered to be unacceptable (Catts et al., 2009; Compton et al., 2006: Jenkins, 2003).

Cut Score Associated with the 40th Percentile

Identifying a cut score associated with a predetermined percentile is a common approach to educational decision-making used in schools (Ochs et al., 2020). The 40th percentile was chosen based on previous research and recommendations from test developers (Clemens et al., 2015; Ochs et al., 2020). Table 7 reports the results of these analyses. See Appendix B for Tables B1, B2, and B3 for the resulting contingency matrices. When calculating the diagnostic accuracy statistics associated with the 40th percentile for each computer-adaptive universal screener, none of the screeners met the optimal sensitivity threshold of .90. Istation (cut score = 230) had the lowest sensitivity at the 40th percentile, with only 49% of students who were identified as being at-risk on the screener later failing the statewide achievement test. The sensitivity for Star Reading (cut score = 467) was slightly stronger at .58, meaning that 58% of students who failed the statewide achievement test were accurately identified as being at-risk by the universal screener. MAP (cut score = 185) had the highest level of sensitivity with 62% of students who failed the statewide achievement test having been identified in Grade 2 as being at-risk. The cut scores associated with the 40th percentile for all measures resulted in a high number of false negatives, students who were categorized as not at-risk who later failed

the statewide achievement measure. Also, adopting a cut score that represented the 40th percentile, resulted in hit rates, or percentage of cases accurately classified, ranging from 66% to 76% across the three universal screening instruments.

Despite the low sensitivity values, all three computer-adaptive universal screeners met the minimum standard of .70 for specificity. Specificity refers to the proportion of students who were not at-risk based on their performance on the screener who went on to pass the statewide achievement test out of the total number of students who passed the state test. For example, Istation and MAP had specificity values above .90, which is considered in the excellent range. In addition, the specificity of Star Reading (.78) was within the acceptable range. The adoption of cut sores representing performance at the 40th percentile on each of the universal screening instruments resulted in higher positive predictive power but lower levels of negative predictive power. Positive predictive power is the probability that students identified as at risk on the screener also failed the statewide achievement test. Negative predictive power is the probability that students identified as not at risk on the screener will pass the statewide achievement test. Istation had a positive predictive power, of .94. which means a student who is identified as at-risk on the screener has a 94% probability of failing the statewide achievement test. MAP had a positive predictive power of .90, which is considered excellent. Also, the positive predictive power of Star Reading fell into the acceptable range (.78). In contrast, the negative predictive power, fell into the unacceptable (<.70) range for all screeners. Unlike sensitivity and specificity, positive and negative predictive power are impacted by the prevalence rate of failure in the sample (i.e., not passing the ACT Aspire), which was > 50% for all three samples.

Table 7

Diagnostic Accuracy for Each Screener by Analysis

Measure and Analysis	Cut Score	Sensitivity	Specificity	PPP	NPP	OCC
40th Percentile						
Istation	230	.49	.96	.94	.58	.69
MAP	185	.62	.92	.90	.68	.76
Star Reading	467	.58	.78	.78	.57	.66
.90 Sensitivity						
Istation	245.3	.90	.65	.77	.83	.79
MAP	195.5	.90	.62	.74	.85	.78
Star Reading	554.5	.90	.45	.69	.76	.71
Max SE and SP						
Istation	243.4	.86	.70	.80	.79	.85
MAP	193.5	.87	.70	.77	.83	.79
Star Reading	490.5	.67	.70	.76	.61	.68

Note. SE = sensitivity*;* SP = specificity; PPP = Positive predictive power, NPP = Negative predictive power, OCC = Overall Correct Classification

Cut Score Associated with Sensitivity Set to .90

Setting sensitivity at .90 resulted in decreased specificity and positive predictive power, but higher negative predictive power for all screeners. Table 7 displays the results. See Appendix C for Tables C1, C2, and C3 for the resulting contingency matrices

When the sensitivity was set at .90, specificity values fell below the acceptable range for all three screeners (range = .45 to .65). This level of sensitivity is desirable for accurately identifying nine out of ten students who are truly at-risk for failure on the state achievement test. Istation (cut score = 245.3) had a specificity value of .65, the highest of the three measures, but below the acceptable level for diagnostic accuracy. MAP (cut score = 195.5) had a specificity value of .62, and the Star Reading (cut score = 554.5) test had the lowest specificity value at .45. When the sensitivity is set at .90, all three measures had higher numbers of false positives, which can result in students who are truly not at risk being given unnecessary interventions which can strain a school's resources.

Setting the specificity at .90 decreased positive predictive power but increased negative predictive power. Although positive predictive power, decreased, it remained within the acceptable range for both Istation (.77) and MAP (.74) but fell within the unacceptable range for Star Reading (.69). These results show that students identified as at risk on the screener have approximately a three out of four chance of failing the statewide achievement test. While positive predictive power, decreased somewhat, the negative predictive power increased with all values being within the acceptable or good ranges. MAP had the highest level of negative predictive power, (.85). Istation had a negative predictive power of .83, and Star Reading had a negative predictive power of

.76. These values indicate that when sensitivity is set at .90, there is a higher probability that a student who is not at risk on the screener will pass the statewide achievement test.

Cut Score Associated with Maximized Sensitivity and Specificity

To maximize both sensitivity and specificity, a cut score was identified using .80 for sensitivity and .70 for specificity as suggested by Kilgus and colleagues in their 2014 meta-analysis on the diagnostic accuracy of universal screeners. Ochs and her colleagues (2020) also used ROC curves to identify cut scores that maximized sensitivity and specificity with a minimum specificity of .70. Table 7 displays the results of the analyses. See Appendix D for Tables D1, D2, and D3 for the resulting contingency matrices

The cut scores associated with maximized sensitivity and specificity resulted in overall correct classification rate statistics between .68 and .85. MAP (cut score = 193.5) had the highest levels of sensitivity (.87) and specificity (.70). Istation (cut score = 243.5) had the highest hit rate (.85) while maintaining an acceptable level of sensitivity (.86) and an acceptable level of SP (.70). Star Reading (cut score = 490.5) had an acceptable level of specificity (.70) but failed to reach an acceptable level of sensitivity (.67). Sensitivity levels above .67 were associated with levels of specificity below .70.

For all three computer-adaptive universal screeners, maximized sensitivity and specificity resulted in adequate levels of positive predictive power, but varying levels of negative predictive power. Istation had the highest level of positive predictive power (.80) with an acceptable level of negative predictive power (.79). MAP had a positive predictive power of .77, and had the highest level of negative predictive power (.83). Star Reading had the lowest levels of positive predictive power (.76) and negative predictive power (.61).

CHAPTER V: DISCUSSION

Universal screening is the first step in identifying students who may be at-risk for not achieving reading proficiency as measured by statewide achievement tests (Klingbeil et al., 2015). In order for accurate and valid decisions to be made, universal screeners must be efficient and technically accurate. The purpose of this study was to compare the diagnostic accuracy of three computer-adaptive universal screeners for predicting performance on a statewide achievement test. To date, no published studies have compared the predictive validity and diagnostic accuracy of three computer-adaptive screeners to one another. Overall, the findings support the use of computer-adaptive universal screeners as strong predictors of statewide measures of reading. However, the results of this study point out a number of relevant findings regarding the setting of cut scores and the subsequent impact on the accuracy of universal screeners.

Predictive Utility of Computer-Adaptive Universal Screeners

Each of the computer-adaptive universal screeners was found to be a strong predictor of the statewide achievement test, ACT Aspire. This finding supports previous research that has demonstrated the strength of relationship between one or more of these computer-adaptive screeners and various statewide achievement tests (Ball & O'Connor, 2016; Clemens et al., 2015; Klingbeil et al., 2018; Ochs et al., 2020; Thomas & January, 2021; Van Norman et al., 2017). When compared to one another, Istation and MAP had stronger correlations to the ACT Aspire than Star Reading.

Computer-adaptive screeners offer schools an alternative to the traditional curriculum-base measures such as oral reading fluency that are commonly used for universal screening. They also offer an approach to screening all students that can limit

the amount of time necessary for test administration and scoring while expanding the domains of skills that can be assessed within a single administration. The efficiency of computer-adaptive screeners as well as their strong relationships to performance on statewide achievement measures have prompted the increased attention they have received in the research literature.

Accuracy of Screening Measures

However, knowing that universal screeners are predictive of statewide achievement tests may be necessary, but it is also insufficient. Universal screeners must also be accurate predictors of a gold standard outcome measure. In this study, the ability of these assessments to predict pass or fail on a statewide achievement test was adopted as the criterion. The area under the curve is considered to be a statistic of overall classification accuracy with values greater than .80 considered to be good (Catts et al., 2009). Istation and MAP met this standard while Star Reading fell below it. Unlike previous literature on diagnostic accuracy, we analyzed the differences between areas under the curves for significance. We found that although Istation and MAP were not reliably different from one another, they both were significantly different from Star Reading. This finding would indicate that Istation and MAP were more accurate classifiers of students' pass/fail performance on the ACT Aspire.

This study has also highlighted that despite strong predictive validity, computer adaptive screeners have varying degrees of accuracy that often fail to meet the recommended levels of sensitivity and specificity, another finding supported in the research literature (Clemens et al., 2015; Kent et al., 2019; Klingbeil et al., 2015; Klingbeil, et al., 2018; Thomas & January, 2021). When using a cut score associated with

a common school approach to setting cut scores for risk, namely the 40th percentile, none of the screeners met the .90 standard for sensitivity. Practically speaking, these screeners would correctly identify only five to six out of every ten students at risk for failure on the ACT Aspire. However, with the cut score at the 40th percentile, Istation and MAP had excellent specificity which means that nine out of ten students who are not at risk would be accurately identified. For Star Reading, approximately eight out of ten students not at risk would be accurately identified. These results show that cut scores set to the 40th percentile on these measures may fail to identify almost half of students in need of intervention support who are on a trajectory to fail the state test.

When sensitivity was set to .90, a standard suggested in the literature by others that ensures that nine of out ten students who are at risk are accurately identified, the overall accuracy of each measure increased to above 70%, but the specificity dropped below acceptable levels. The cut scores associated with the sensitivity level of .90 would allow 90% of students who were truly at risk for failure on the ACT Aspire to be identified accurately. Yet, more students who are not at risk may be inaccurately identified as being at risk which can result in available school resources being deployed unnecessarily.

When sensitivity and specificity were maximized, the highest overall correct classification rate of 85% was achieved for all analyses. Istation maintained a sensitivity level of .86 while also holding specificity to .70, the recommended acceptable level while accurately identifying 85% of students in the sample. MAP also resulted in its highest hit rate of 79% while maintaining a sensitivity level of .87 and a specificity level of .70. Only Star Reading did not achieve an acceptable rate of sensitivity when sensitivity and

specificity were maximized. It also had a hit rate that was lower than the hit rate when sensitivity was set to .90. These results indicate that the method for choosing cut scores may depend on the test being used and the population for whom the test is administered. In our study, maximizing sensitivity and specificity was the best approach for identifying cut scores with appropriate levels of accuracy for two of the three measures.

Sensitivity and specificity are traditionally believed to be aspects of the tests themselves. However, more recent consensus from the medical field suggests that these indices are not fixed properties of the measures (Bossuyt et al., 2015). Instead, these indices may vary across setting, context, and sample which could account for why different methods for establishing cut scores were found to be the better approach for the different computer-adaptive universal screeners.

The positive predictive power and negative predictive power are sample-based indices that some argue, should be more closely considered when choosing a universal screener or making decisions with its data (Trevathan, 2017). The current results showed that when sensitivity was high, then negative predictive power was also high. When sensitivity was low, as demonstrated by the results of the cut scores set at the 40th percentile, positive predictive power was high. For Istation and MAP, all values were acceptable when sensitivity and specificity were maximized. Star Reading did not achieve the same results.

One possible explanation for the differences in diagnostic accuracy and the ways cut scores are set could lie in the design and content of each measure. Istation and MAP have a suite of assessments designed to assess skills across a wide range of grade levels, including early literacy skills. Star Reading focuses solely on the higher order reading

skills of vocabulary, comprehension, and literacy elements and characteristics that align to state and national curriculum standards. It is possible that Istation and MAP are more accurate predictors because they are able to assess a wider range of reading skills that allows a more accurate prediction of performance on the state test.

Practically, Istation and MAP demonstrated adequate diagnostic accuracy despite failing to meet the recommended sensitivity level of .90. Although all three computer-adaptive universal screeners were strong predictors of the ACT Aspire, they demonstrated differences in accuracy, an important aspect of screeners to be considered. The three different cut scores for each screener resulted in varying degrees of accuracy. Therefore, schools should carefully consider both the benefits and drawbacks of the screener, including the accuracy of the results.

Implications

In addition to the predictive validity and diagnostic accuracy of the computer-adaptive universal screeners, the findings of this study provide implications for practioners seeking to enhance the effectiveness universal screening within RTI. The primary goal of RTI is to change the trajectories of students who are identified as being at-risk for future reading problems (Kent et al., 2019). Yet, so often schools' available resources to intervene appropriately are limited. Therefore, having a universal screener that is accurate is of utmost importance.

A direct route approach to screening involves using a single screener on which to base decisions about supports. When using a direct route approach, schools must carefully select a universal screener. Diagnostic accuracy statistics can support schools in choosing an appropriate screener. Only when sensitivity and specificity were maximized

would Istation and MAP be close to acceptable for use in a direct route approach to universal screening with sensitivity levels slightly below .90 and specificity levels of .70. Based on the criteria set by the National Center for Intensive Intervention (n.d.), both screeners show partially convincing evidence for adequacy as single universal screening measures. Star Reading failed to achieve acceptable levels of both sensitivity and specificity with any of the three analyses in this study.

Decision makers must consider the practical utility when choosing a universal screener and balance their needs with the trade-offs (Petscher et al., 2011). Catts et al. (2015) argues that the goal of RTI is to identify and intervene for as many students as possible. Vaughn and Fletcher (2020) agreed with the notion that overidentification is preferable to under-identification because of the potential repercussions for students who fail to get the intervention vital to their success as readers. For this to happen, a sensitivity level of .90 is necessary. However, a high level of sensitivity results in more false positives that can stretch the already limited resources of some schools. According to the results of this study, maximizing both sensitivity and specificity can lead to more accurate classifications overall, but with lower levels of sensitivity and specificity. Establishing a local cut score that allows for high sensitivity and acceptable specificity could be a more agreeable first step for most schools.

Another approach schools might consider is using multiple measures for universal screening or including additional sources of data. Although this study focused on the use of single universal screeners administered to independent groups, other studies have evaluated a screening approach that uses multiple screening measures to reduce the number of false positives from a single screener or includes additional information such

as previous year's test scores. Klingbeil and colleagues (2015) found that using multiple assessments including oral reading fluency led to more accurate decision-making, a finding also supported by Compton et al. (2006) and Johnson et al. (2010). Ball and O'Connor (2016) found that multi-screener approaches demonstrated strong accuracy over any single measure. Van Norman and his colleagues (2017) found that adding the previous year's test scores to MAP scores resulted in better predictions of state test performance. The benefits of more accurate identification must be considered against the instructional time needed to administer multiple measures. Including already available information is one way to increase the accuracy of a single screener without increasing the amount of time needed for more assessments.

A gated approach to screening is an alternative to administering multiple assessments to each student in a single administration and uses a multilevel, diagnostic framework towards screening. In this approach, the initial screener cut score is set to identify as many students who are at-risk for failure on the state test as possible. For these purposes, the recommended sensitivity level of .90 would be appropriate. Then, only students who are identified as being at-risk are administered additional assessments to further distinguish between true positives and false positives. These additional measures should increase in intensity and accuracy to increase the accuracy of the screening (Walker et al., 2014). VanMeveren and her colleagues (2020) found that the diagnostic accuracy of universal screening was improved when including the previous year's test scores for fourth- and fifth-graders and administering an oral reading fluency measure to students who were identified as at-risk on MAP. Gated approaches offer an alternative to relying solely on a single measure. However, further research is needed to identify how to

increase the diagnostic accuracy of a universal screening approach that minimizes the resources necessary from administration and the amount of time necessary for testing.

Limitations and Future Directions

Despite the moderate to strong evidence supporting the diagnostic accuracy of computer-adaptive screeners, we note the following limitations with this study. First, the data originated from a single state which limits the generalizability of the findings. Although the analytic sample closely mirrored the state population of students in second grade in the 2017-2018, consideration should be given to the characteristics of the students when attempting to generalize the results. As discussed previously, diagnostic accuracy metrics are dependent on the setting, context, and sample on which the analyses are conducted. Although the demographics of the analytic sample in this study were comparable across the universal screener measures, the generalizability of the findings is restricted.

The findings of this study are also limited by the uneven sample sizes across the three computer-adaptive tests. For example, the analytic sample for Star Reading had a much lower percentage of limited English proficient students (3.7%) compared to MAP (14%). However, having a lower number of students for whom English is not a first language would typically result in stronger results, but that was not the case in this study. Star Reading did have a higher percentage of students on free or reduced lunch (70.6%) as compared to MAP (59.0%) and Istation (65.1%). Free or reduced lunch status has been shown to impact student achievement and therefore could account for the performance of Star Reading in this study. Star Reading also had more schools that scored in the D and F category for school ratings (27.4%) then Istation (5.9%). The context in which the Star

Reading assessment is being administered could account for the results reported in this study. Future research that accounts for school and student demographic characteristics could address the limitations of this study.

Another limitation is the number of students who were dropped from the analytic sample due to missing screener or ACT Aspire scores. Without the test scores, it is not possible to know how the sample sizes for each screener may have been impacted by the exclusion of those students.

This study is also limited by the focus on predicting student achievement outcomes in grade 3 and is therefore not generalizable to other grades. It is possible that Istation and MAP showed better predictive capability because of their inclusion of print skills such as phonemic awareness and decoding which are more predictive at grade 3 whereas Star Reading assessed higher-order comprehension skills. Research has shown that as students move through higher grades, the influence of decoding on reading comprehension decreases (Language and Reading Research Consortium, 2015). The influences that are most predictive of reading comprehension shift across the grades and may impact the predictive abilities of the measures. Therefore, the findings from this study are constrained to the grade levels included.

Another limitation of this study is the elapsed time between the universal screening assessment in the spring of second grade and the state reading assessment administered in the spring of third grade which could introduce confounding variables such as the impact of possible summer reading loss and student transition between schools. Efforts were made to ensure students could be tracked across schools, especially for those students who attended a different school for third grade in order to retain them

in the analytic sample. However, bridging from one school year to the next school year resulted in some students being dropped from the analytic sample due to incomplete data. However, the time elapsed is also a strength in that this study shows that risk can be identified early so that intervention can start before students must take the state test.

Despite the statewide achievement test used in the study, the ACT Aspire, being a nationally normed and available test, the results are limited to the universal screeners' accuracy in relation to performance on this measure. Still, this study does extend the literature by using a measure that is not necessarily limited to a single state as most state achievement tests are designed around a single state's curriculum standards. Moreover, the ACT Aspire is used in multiple states.

Finally, our data set did not include information regarding the instruction or intervention students received between the test administrations. Within a MTSS framework, students classified as at risk would have received intervention in addition to classroom instruction. It is possible that false positives were due to intervention being implemented between the screener and state test. However, this information was beyond the scope of our extant data. Additionally, the results observed for all three screeners seem to indicate that the instruction and intervention may not be adequately meeting the needs of students who are identified as being at risk. If the instruction and intervention were appropriate, then we would expect to have seen more false positives, meaning students were identified as at risk on the screener in grade 2, but went on to pass the state test in grade 3. The context of RTI can and should impact the diagnostic accuracy of the universal screeners.

Additional research is needed to replicate and extend these finding to other computer-adaptive universal screeners and corresponding state tests. Future research should include the impact of student and school characteristics on the diagnostic accuracy of computer-adaptive measures since the context and setting impact the accuracy metrics examined. Few studies were found that included student characteristics in their analyses of the diagnostic accuracy of universal screeners. Understanding how school characteristics such as overall performance as indicated by letter grade designations or free and reduced lunch rate and student characteristics such as special education designation or economically disadvantaged status impact the diagnostic accuracy of a universal screener can lead to better decision making about criteria being set. It can also support the evaluation of the effectiveness of the instruction and intervention being provided to students.

Also, continued research that seeks to enhance the balance of sensitivity, specificity, positive predictive power, negative predictive power, and overall accuracy is needed. Research should explore earlier grade levels in which risk can be accurately identified to allow more time to intervene and change the trajectories of students with the ultimate goal being student success. Finally, research should establish ways for local schools and districts to determine the diagnostic accuracy of their universal screeners that do not involve complicated statistical analyses to which they may not have access. Creating the contingency matrices such as those contained in Appendices B, C, and D can be a beginning point for this work.

Conclusions

In conclusion, computer-adaptive screeners have demonstrated strong predictive validity with varying degrees of diagnostic accuracy. These results indicate that establishing local cut scores that maximize the accurate identification of students who are at-risk produce more accurate results than following the recommended national cut score. It is also critical that educators become informed practioners on how to use data for planning instruction and intervention. With the crucial decisions being made based on the outcomes of universal screeners, it is important that the tools being used provide the most accurate results.

REFERENCES

ACT, Inc. (2020). *ACT aspire summative technical manual*. ACT, Inc. https://actinc.my.salesforce.com/sfc/p/#300000000Wu5/a/4v0000005fHp/SLZ26XzhfmI8ibKP_Ca5G94_T3HuveFbNgFmfcRaHoY

ACT, Inc. (2019). ACT Aspire Summative Report, AR Grade 3. file:///Users/susanporter/Dropbox/MTSU%201/Dissertation/AR%20Documents/ACT%20Aspire%202019%203rd%20Gr%20Report.pdf

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*, 34-53. https://doi.org/10.1007/s11881-018-00171-5

Al Otaiba, S., Folsom, J. S., Schatschneider, C., Wanzek, J., Greulich, L., Meadows, J., Li, Z., & Connor, C. M. (2011). Predicting first-grade reading performance from kindergarten response to tier 1 instruction. *Exceptional Children, 77*(4), 453-470. https://doi.org/10.1177/001440291107700405

Al Otaiba, S., Kim, Y., Wanzek, J., Petscher, Y., & Wagner, R. K. (2014). Long-term effects of first-grade multitier intervention. *Journal of Research on Educational Effectiveness, 7*, 250-267. https://doi.org/10.1080/19345747.2014.906692

Al Otaiba, S., & Petscher, Y. (2020). Identifying and serving students with learning disabilities, including dyslexia, in the context of multitiered supports and response to intervention. *Journal of Learning Disabilities, 53*(5), 327-331. https://doi.org/10.1177/0022219420943691

Anne E. Casey Foundation. (2010). *Early warning! Why reading by the end of third grade matters*. https://www.aecf.org/resources/early-warning-why-reading-by-the-end-of-third-grade-matters

Armstrong, N., & Eborall, H. (2012). The sociology of medical screening: Past, present and future. *Sociology of Health & Illness, 34*(2), 161-176. https://doi.org/10.1111/j.1467-9566.2011.01441.x

Bailey, T. R. (2019, September 20) Is mtss the new rti? Depends on where you live. *Center on Multi-Tiered System of Supports.* https://mtss4success.org/blog/mtss-new-rti-depends-where-you-live

Ball, C. R., & O'Connor, E. O. (2016). Predictive utility and classification accuracy of oral reading fluency and the measures of academic progress for the Wisconsin knowledge and concepts exam. *Assessment for Effective Intervention, 41*(4), 195-208. https://doi.org/10.1177/1534508415620107

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., Lijmer, J. G., Moher, D., Rennie, D., de Vet, H. C. W., Kressel, H. Y., Rifai, N., Golub, R. M., Altman, D. G., Hooft, L., Korevaar, D. A., Cohen, J. F., & STARD Group (2015). STARD 2015: A updated list of essential items for reporting diagnostic accuracy studies. *Clinical Chemistry, 61(12), 1446 - 1452*. https://doi.org/10.1373/clinchem.2015.246280

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

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

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. http://dx.doi.org/10.17105/SPR44-1.76-97

Clemens, N. H., Oslund, E., Kwok, O., Fogarty, M., Simmons, D., & Davis, J. L. (2019). Skill moderators of the effects of a reading comprehension intervention. *Exceptional Children, 82*(2), 197-211. https://doi.org/10.1177/0014402918787339

Clemens, N. H., Oslund, E. L., Simmons, L. E., & Simmons, D. (2014). Assessing spelling in kindergarten: Further comparison of scoring metrics and their relation to reading skills. *Journal of School Psychology, 52*, 49-61. https://doi.org/10.1016/j.jsp.2013.12.005

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (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

Compton, D. L., Gilbert, J. K., Jenkins, J. R., Fuchs, D., Fuchs, L. S., Cho, E., Barquero, L. A., & Bouton, B. (2012). Accelerating chronically unresponsive children to tier 3 instruction: What level of data is necessary to ensure selection accuracy?

Journal of Learning Disabilities, 45(3), 204-216. https://doi.org/10.1177/0022219412442151

Cook, M. A., & Ross, S. M. (2020). PARCC predictability study – 3rd grade. Johns Hopkins University. https://www.istation.com/Content/downloads/studies/PARCCPredictabilityAnalyses.pdf

Deno, S. L. (2003). Curriculum-based measures: Development and perspectives. *Assessment for Effective Intervention, 28*(3-4), 3-12. https://doi.org/10.1177/073724770302800302

Deno, S. (n.d.). Ongoing student assessment. RTI Action Network. Retrieved from http://www.rtinetwork.org/essential/assessment/ongoingassessment.

Ehri, L. C. (2004). Teaching phonemic awareness and phonics: An explanation of the national reading panel meta-analysis. In P. McCardle & V. Chhabra (Eds.), *The voice of evidence in reading research* (pp. 153-186). Paul H. Brookes Publishing Company.

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

Elementary & Secondary Education Act of 1965, 20 U.S.C. *§* 6301 (1965). https://www2.ed.gov/documents/essa-act-of-1965.pdf

Every Student Succeeds Act, 20 U.S.C. *§* 6301 *114* (2015). https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf

Fletcher, J. M., Francis, D. J., Morris, R. D., & Lyon, G. R. (2005). Evidence-based assessments of learning disabilities in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34*(3), 506-522. https://doi.org/10.1207/s15374424jccp3403_7

Fletcher, J. M., Francis, D. J., Foorman, B. R., & Schatschneider, C. (2021). Early detection of dyslexia risks: Development of brief, teacher-administered screens. *Learning Disability Quarterly, 44*(3), 145-157. https://doi.org/10.1177/0731948720931870

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2019). *Learning Disabilities: From Identification to Intervention* (2nd ed.). Guilford.

Francis, D. J., Fletcher, J. M., Stuebing, K. K., Lyon, G. R., Shaywitz, B. A., & Shaywitz, S. E. (2005). Psychometric approaches to identification of LD: IQ and achievement scores are not sufficient. *Journal of Learning Disabilities, 38*(2), 98-108. https://doi.org/10.1177/00222194050380020101

Fuchs, D., & Fuchs, L. S. (2017). Critique of the National Evaluation of Response to Intervention: A Case for Simpler Frameworks. Exceptional Children, 83(3), 255–268. https://doi.org/10.1177/0014402917693580

Fuchs, D., & Fuchs, L., S. (2009). On the importance of a unified model of responsiveness to intervention. *Child Development Perspectives, 3*(1), 41-43. https://doi.org/10.1111/j.1750-8606.2008.00074.x

Fuchs, D., Fuchs, L. S., & Stecker, P. M. (2010). The “blurring” of special education in a new continuum of general education placements and services. *Exceptional Children, 76*(3), 301-323. https://doi.org/10.1177/001440291007600304

Glover T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45*, 117-135. https://doi.org/10.1016/j.jsp.2006.05.005

Good, R. H., & Kaminski, R. A. (2020) Dynamic Indicators of basic early literacy skills (8th ed.). Institute for the Development of Educational Achievement.

Graham, S., & Santangelo, T. (2014). Does spelling instruction make students better spellers, readers, and writers? A meta-analytic review. *Reading and Writing Quarterly, 27*, 1703-1743. https://doi.org/10.1007/s11145-014-9517-0

Hajian-Tilaki, K. (2013). Receiver operating characteristic curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine, 4*(2), 627-635.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29-36. https://doi.org/10.1148/radiology.143.1.7063747

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under the receiver operating characteristic curves derived from same cases. *Radiology, 148*, 839-843.

Hasbrouck, J. & Tindal, G. (2017). *An update to compiled ORF norms* (Technical Report No. 1702). University of Oregon. https://files.eric.ed.gov/fulltext/ED594994.pdf

Hosp, J. L., & Ardoin, S. P. (2008). Assessment for instructional planning. *Assessment for Effective Intervention, 33*(2). 69-77. https://doi.org/10.1177/1534508407311428

Hosp, J. L., Hosp, M. A., & Dole, J. K. (2011). Potential bias in predictive validity of universal screening measures across disaggregation subgroups. *School*

Psychology Review, 40(1), 108-131. https://doi.org/10.1080/02796015.2011.12087731

Hutt, E., & Schneider, J. (2018). A history of achievement testing in the united states or: Explaining the persistence of inadequacy. *Teachers College Record, 120*(11). https://www.tcrecord.org/Content.asp?ContentId=22452

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 (2004).

January, S. 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

Jeon, H., Wall, S. M., Peterson, C. A., Luze, G. J., & Swanson, M. E. (2018). Using early indicators of academic risk to predict academic skills and socioemotional functioning at age 10. *Journal of Educational Psychology, 110*(4), 483-501. https://doi.org/10.1037/edu0000230

Johnson, E. S., Jenkins, J. R., & Petscher, Y. (2010). Improving the accuracy of a direct route screening process. *Assessment for Effective Intervention, 35*(3), 131 – 140. https://doi.org/10.1177/1534508409348375

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

Kane, M. T. (2013). The argument-based approach to validation. *School Psychology Review, 42*, 448-457.

Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review, 37*(3), 374-390. https://doi.org/10.1080/02796015.2008.12087884

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 Instruction, 44*(3), 160-172. https://doi.org/10.1177/1534508418758371

Kettler, R. J., & Albers, C. A. (2013). Predictive validity of curriculum-based measurement and teacher ratings of academic achievement. *Journal of School Psychology, 51*, 499-515. https://doi.org/10.1016/j.jsp.2013.02.004

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*, 377-405. https://doi.org/10.1016/j.jsp.2014.06.002

King, K. R., Lembke, E. S., & Reinke, W. M. (2016). Using latent class analysis to identify academic and behavioral risk status in elementary students. *School Psychology Quarterly, 31*(1), 43-57. https://doi.org/10.1037/spq0000111

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

Klingbeil, D. A., Van Norman, E. R., Nelson, P. M., & Birr, C. (2018). Evaluating state procedures across changes to the statewide achievement test. Assessment for Effective Intervention, 44(1), 17-31. https://doi.org/10.1177/1534508417747390

Kuhn, M., & Johnson, K. (2016). *Applied predictive modeling*. Springer.

Lam, E. A., & McMaster, K. L. (2014). Predictors of responsiveness to early intervention: A 10 year update. *Learning Disability Quarterly, 37*(3), 134-147. https://doi.org/10.1177/0731948714529772

Language and Reading Research Consortium (2015). Learning to read: Should we keep things simple? Reading Research Quarterly, 50(2), 151-169. https://doi.org/10.1002/rrq.99

Leonard, K. M., Coyne, M. D., Oldham, A. C., Burns, D., & Gillis, M. B. (2019). Implementing mtss in beginning reading: Tools and systems to support schools and teachers. Learning Disabilities Research & Practice, 34(2), 110-117. https://doi.org/10.1111/ldrp.12192

Lyon, G. R. (1998). *Overview of reading and literacy initiatives*. (ED444128). ERIC. https://eric.ed.gov/?id=ED444128

Lyon, G. R., Shaywitz, S. E., Shaywitz, B. A., Chhabra, V. (2005). Evidence-based reading policy in the united states: How scientific research informs instructional practice. *Brookings Papers on Education Policy, 8*, 209-250. http://www.jstor.org/stable/20062559

Mathes, P. (2009). *Istation's indicators of progress early reading reliability and validity evidence*. Istation. https://www.istation.com/Content/downloads/studies/isip_rr.pdf

Mathes, P., Torgesen, J., & Herron, J. (2016). *Computer adaptive testing system for continuous progress monitoring for reading growth for students pre-k through grade 3*. Istation. https://www.istation.com/Content/downloads/studies/er_technical_report.pdf

McKenna, M. C., & Stahl, K. A. D. (2009). *Assessment for reading instruction* (2nd ed.). The Guildford Press.

Morken, F., Jones, L. O., & Helland, W. A. (2021). Disorders of language and literacy in the prison population: A scoping review. *Education Sciences, 11*(2), 77. https://doi.org/10.3390/educsci11020077

National Center for Intensive Intervention, (n.d.). Academic screening tools chart. https://intensiveintervention.org/tools-charts/overview?_ga=2.142849586.607777730.1662837233-684324072.1660011889

Nese, J. F. T., Park, B. J., Alonzo, J., & Tindal, G. (2011). Applied curriculum-based measurement as a predictor of high-stakes assessment. *The Elementary School Journal, 111*(4), 608-624. https://doi.org/10.1086/659034

Nietzel, M. T. (2020). Low literacy levels among U.S. adults could be costing the economy $2.2 trillion a year. *Forbes*. https://www.forbes.com/sites/michaeltnietzel/2020/09/09/low-literacy-levels-among-us-adults-could-be-costing-the-economy-22-trillion-a-year/?sh=69e2cd054c90

No Child Left Behind Act of 2001, Pub. L. No. 107-110 (2001). https://www.congress.gov/bill/107th-congress/house-bill/1/text

NWEA. (2019). *MAP growth technical report*. NWEA. https://www.nwea.org/content/uploads/2021/11/MAP-Growth-Technical-Report-2019_NWEA.pdf

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 Instruction, 45*(3), 210-225. https://doi.org/10.1177/1534508418796271

Odegard, T. N., Farris, E. A., Middleton, A. E., Oslund, E., & Rimrodt-Frierson, S. (2020). Characteristics of students identified with dyslexia within the context of state legislation. *Journal of Learning Disabilities, 53*(5), 366-379. https://doi.org/10.1177/0022219420914551

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, 53*(6), 243-255. https://doi.org/10.1177/0022219420972187

Patarapichayatham, C., & Locke, V. N. (2020). *Linking the ACT aspire assessments to ISIP reading and math.* Istation. https://www.istation.com/Content/downloads/studies/ACT_Aspire.pdf

Paul, C. A. (2016). *Elementary and Secondary Education Act of 1965*. Social Welfare History Project. https://socialwelfare.library.vcu.edu/programs/education/elementary-and-secondary-education-act-of-1965/

Peng, P., Fuchs, D., Fuchs, L. S., Elleman, A. M., Kearns, D. M., Gilbert, J. K., Compton, D. L., Cho, E., & Patton, S. (2019). A longitudinal analysis of the

trajectories and predictors of word reading and reading comprehension development among at-risk readers. *Journal of Learning Disabilities, 52*(3), 195-208. https://doi.org/10.1177/0022219418809080

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

Petscher, Y., Fien, H., Stanley, C., Gearin, B., Gaab, N., Fletcher, J. M., & Johnson, E. (2019a). *Screening for dyslexia*. U.S. Department of Education, Office of Elementary and Secondary Education, Office of Special Education Programs, National Center on Improving Literacy. https://improvingliteracy.org/whitepaper/screening-dyslexia

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

Petscher, Y., Solari, E. J., & Catts, H. W. (2019b). Conditional longitudinal relations of elementary literacy skills to high school reading comprehension. *Journal of Learning Disabilities, 52*(4), 324-336. https://doi.org/10.1177/0022219419851757

Plato. (1968). *The Republic*. (A. Bloom, Trans.). Basic Books.

Reese, W. J. (2013). *Testing wars in the public schools: A forgotten history*. Harvard University Press.

Renaissance Learning. (2022). Star assessments for reading technical manual. Renaissance Learning. https://help.renaissance.com/US/PDF/SR/SRRPTechnicalManual.pdf

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

Schatschneider, C., Wagner, R. K., & Crawford, E. C. (2008). The importance of measuring growth in response to intervention models: Testing a core assumption. *Learning and Individual Differences, 18*, 308-315. https://doi.org/10.1016/j.lindif.2008.04.005

Shapiro, E. S., & Gebhardt, S. N. (2012). Comparing computer-adaptive and curriculum-based measurement methods of assessment. *School Psychology Review, 41*(3), 295-305. https://doi.org/10.1080/02796015.2012.12087510

Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-based measures and performance on the 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

Shin, J., & McMaster, K. (2019). Relations between CBM (oral reading fluency 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

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

Silva, M., & Cain, K. (2015). The relations between lower and higher level of comprehension skills and their role in prediction of early reading comprehension. *Journal of Educational Psychology, 107*(2), 321-331.

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,* 288-308. https://doi.org/10.1111/bjep.12150

Snow, C., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children.* National Academy Press.

Speece, D. L., Ritchey, K. D., Silverman, R., Schatschneider, C., Walker, C. Y., & Andrusik, K. N. (2010). Identifying children in middle childhood who are at risk for reading problems. *School Psychology Review, 39*(2), 258-276. https://doi.org/10.1080/02796015.2010.12087777

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

Swets, J. A. (1996). *Signal detection theory and roc analysis in psychology and diagnostics: Collected papers*. Lawrence Erlbaum Associates.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better decisions through science. *Scientific American, 283*(4), 82-87. https://www.jstor.org/stable/26058901

Thomas, A. S., & January, S. A. (2021). Evaluating the criterion validity and classification accuracy of universal screening measures in reading. *Assessment for Effective Instruction*, *46*(2), 110-120. https://doi.org/10.1177/1534508419857232

Trevathan, R. (2017). Sensitivity, specificity, and predictive values: Foundations, pliabilities, and pitfalls in research and practice. Frontiers in Public Health, 5, 1 – 7. https://doi.org/10.3389/fpubh.2017.00307

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

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

VanDerHeyden, A. (2010). Use of classification agreement analyses to evaluate RTI implementation. *Theory into Practice, 49*, 281-289. https://doi.org/10.1080/00405841.2010.510721

Vaughn, S., Capin, P., Scammacca, N., Roberts, G., Cirino, P., & Fletcher, J. M. (2020). The critical role of word reading as a predictor of response to intervention. *Journal of Learning Disabilities, 53*(6), 415-427. https://doi.org/10.1177/0022219419891412

Vaughn, S., & Fletcher, J. M. (2021). Identifying and teaching students with significant reading problems. *American Educator, 44*(4), 4 – 11. https://eric.ed.gov/?id=EJ1281906

Vaughn, S., & Fuchs, L. S. (2006). A response to "competing views: A dialogue on response to intervention": Why response to intervention is necessary but not sufficient for identifying students with learning disabilities. *Assessment for Effective Intervention, 32*(1), 58-61. https://doi.org/10.1177/15345084060320010801

Walker, H. M., Small, J. W., Severson, H. H., Seeley, J. R., & Feil, E. G., (2014). Multiple-gating approaches in universal screening within school and community settings. In R. T. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings* (pp. 47-75). American Psychological Association.

Wanzek, J., Al Otaiba, S., & McMaster, K. L. (2020). *Intensive reading interventions for the elementary grades*. The Guilford Press.

Whittaker, M., & Batsche, G. (2019). *Data-based problem solving: Effective implementation of MTSS, RTI, and PBIS*. National Center for Learning Disabilities. https://www.ncld.org/wp-content/uploads/2019/11/Essential-Components-of-Data-Based-Problem-Solving-Approaches.Final_.pdf

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

Zirkel, P. A., & Thomas, L. B. (2010). State laws and guidelines for implementing RTI. *Teaching Exceptional Children, 43*(1), 60-73.

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation, 115*, 654-657. https://doi.org/10.1161/CIRCULATIONAHA.105.594929

APPENDICES

APPENDIX A

ROC Curves for Individual Screeners

Figure A1

ROC Curve for Istation

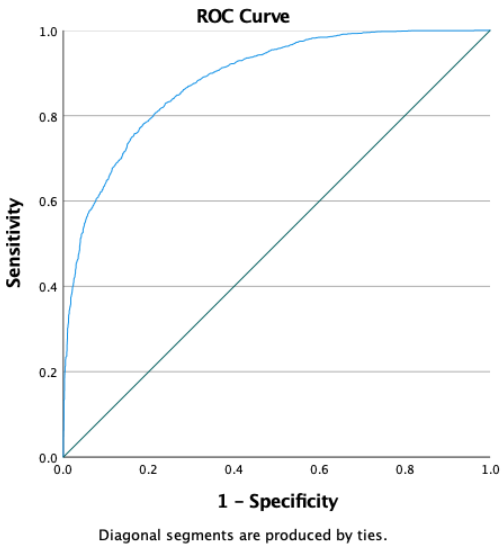


Figure A2

ROC Curve for MAP

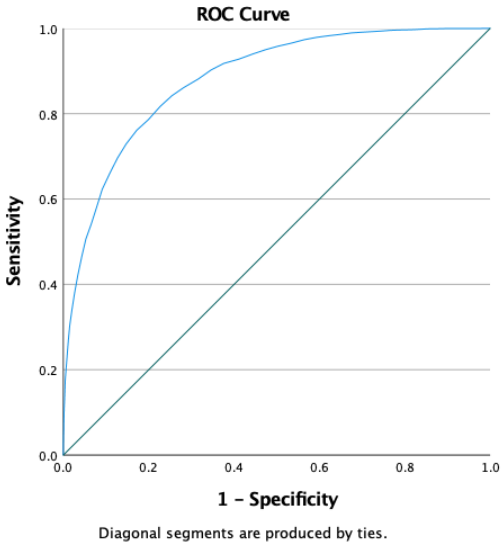
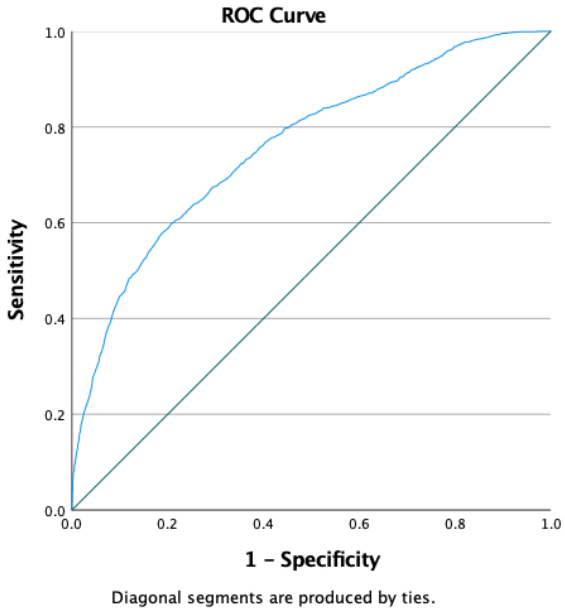


Figure A3

ROC Curve for Star Reading



APPENDIX B

Contingency Matrices for Cut Score Associated with 40th Percentile

Table B1

Contingency Matrix for Cut Score Cut Score Associated with 40th Percentile for Istation

	ACT Aspire	
	Fail	Pass
Fail	1356	81
Pass	1433	1982

Table B2

Contingency Matrix for Cut Score Cut Score Associated with 40th Percentile for MAP

	ACT Aspire	
	Fail	Pass
Fail	4167	470
Pass	2512	5261

Table B3

Contingency Matrix for Cut Score Cut Score Associated with 40th Percentile for Star Reading

	ACT Aspire	
	Fail	Pass
Fail	2020	563
Pass	1470	1977

APPENDIX C

Contingency Matrices for Cut Score Associated with .90 Sensitivity

Table C1

Contingency Matrix for Cut Score Associated with .90 Sensitivity for Istation

		ACT Aspire
	Fail	Pass
Fail	2509	733
Pass	280	1330

Table C2

Contingency Matrix for Cut Score Associated with .90 Sensitivity for MAP

		ACT Aspire
	Fail	Pass
Fail	6070	2160
Pass	609	3571

Table C3

Contingency Matrix for Cut Score Associated with .90 Sensitivity for Star Reading

		ACT Aspire
	Fail	Pass
Fail	3140	1404
Pass	350	1136

APPENDIX D

Contingency Matrices for Cut Score Associated with Maximized Sensitivity (SE) and Specificity (SP)

Table D1

Contingency Matrix for Cut Score Associated with Maximized SE and SP for Istation

	ACT Aspire	
	Fail	Pass
Fail	2408	617
Pass	381	1446

Table D2

Contingency Matrix for Cut Score Associated with Maximized SE and SP for MAP

	ACT Aspire	
	Fail	Pass
Fail	5832	1742
Pass	847	3989

Table D3

Contingency Matrix for Cut Score Associated with Maximized SE and SP for Star Reading

	ACT Aspire	
	Fail	Pass
Fail	2325	755
Pass	1165	1785