Academic Screening in Middle School: How well do AIMSweb Measures of Oral Reading Fluency, and NWEA Measures of Academic Progress, Predict Future Performance on State Exams

Matthew Mitchell

A Thesis Submitted in
Partial Fulfillment of the
Requirements for the Degree of

Educational Specialist
School Psychology

At

The University of Wisconsin-Eau Claire

November, 2019

Graduate Studies

The members of the Committee approve the thesis of

Matthew Mitchell presented on November 21st, 2019

Dr. Mary Beth Tusing

Dr. Jeffrey Goodman

Dr. Barbara Meier

APPROVED: _____

Dean of Graduate Studies

Academic Screening in Middle School: How well do AIMSweb Measures of Oral Reading Fluency, and NWEA Measures of Academic Progress, Predict Future Performance on State Exams

By

Matthew Mitchell

The University of Wisconsin-Eau Claire, 2019
Under the Supervision of Dr. Mary Beth Tusing

The current study examined the diagnostic accuracy of two common screening assessments in reading, Measures of Academic Progress (MAP) and Reading-Curriculum Based Measurement (R-CBM), when used to predict end of the year performance on state tests in 8th grade. The sample consisted of 389 8th grade students enrolled in a school district in the Upper Midwest. Results of this study demonstrate that MAP was the better individual measure when assessing diagnostic accuracy. Further the combination of R-CBM and MAP assessment results did not improve diagnostic accuracy when compared to MAP as a single screening assessment. However, these results suggest using only MAP to screen students in 8th grade may result in many students being misidentified as not at-risk when using publisher recommended cut-off scores. Future middle school research could explore different cut-scores for defining "at-risk" students or more liberal approaches when using a combined screening model.

Mary Beth Tusing 12/6/19

Mary Beth Tusing Date

TABLE OF CONTENTS

	Page
LIST OF TABLES	vi
Chapter	
I. INTRODUCTION	1
Review of Literature	2
Summary of Research	15
Statement of the Problem	15
II. METHODS	17
Participants and Setting	17
Measures	18
Procedure	21
Data Analysis	22
III. RESULTS	24
Descriptive Statistics and Correlations	24
Research Question 1: Single Screener Model	25
Research Question 2: Combined Screener Model	26
IV. DISCUSSION	28
Interpretations	28
Implications for Practice and Future Research	34
Limitations	38
Conclusion	38
REFERENCES	39
APPENDICES	45
A. Definitions of Key Terms in Relation to Academic Screening	45
B. Formulas Used in Hand Calculation Analysis	47

LIST OF TABLES

Table		Page
1.	Sample Characteristics	18
2.	Descriptive Statistics for Reading Screeners and State Test (N = 389)	25
3.	Correlation Between Variables	25
4.	Diagnostic Accuracy of MAP and R-CBM Universal Screening Measures (N=389)	26

Introduction

Current national assessments in the area of reading suggest the majority of U. S. students are not proficient in their ability to read and comprehend material. As a result, educational policies at the federal level such as No Child Left Behind and more recently the Every Student Succeeds Act have shifted education towards a mindset of visible growth and school accountability for the achievement of students of all backgrounds and ages (U.S. Department of Education, 2015). To meet this challenge, a shift has taken place across elementary and primary schools from a skill remediation model to a preventative approach, commonly referred to as Response to Intervention (RtI; Sanger, Friedli, Brunken, Snow, & Ritzman, 2012). In RtI, schools put in place a multi-leveled system of supports to meet the varied needs of students (Johnson & Smith, 2011; Sanger et al., 2012). Further, the National Center on Intensive Intervention indicates data-based decision-making should guide the selection of interventions and intensity of support systems in an RtI framework. School-wide screening is a primary method of data collection that is employed by most schools for RtI decision-making. The intent of this thesis is to add to the current body of research examining the accuracy of screening in the area of reading at the middle school level.

The first chapter of this thesis reviews current best practices when conducting academic screening in a RtI system at the middle school level. Definitions of key terms from the review of research are listed in Appendix A. Terms include: specificity, sensitivity, positive predictive power, negative predictive power, receiver operating characteristic curve, area under the curve, curriculum based measurement, computer adaptive tests, and diagnostic accuracy. Further, feasibility factors and technical

properties are reviewed for two different types of assessments commonly used when screening student's reading skills at the middle school level. Finally, diagnostic accuracy research for individual as well as combined screening techniques are reviewed prior to the statement of research goals.

Review of Literature

Academic Screening in Response to Intervention. Sanger et al. (2012) describes RtI as a three-leveled approach that allows educators to identify students who are "at-risk" for academic and/or behavioral failure to then determine the appropriate level of intervention that is needed for a student. The National Center on Response to Intervention (NCI, 2019, para 2), lists screening, progress monitoring, and multilevel prevention systems as the foundations for making data-based decisions for RtI. This model originated in elementary settings as educators attempted to use academic screening assessments to identify students in need of instructional support beyond the core curriculum (Shapiro, Keller, Lutz, Santoro, & Hintze, 2006). Further, the National Center on Intensive Intervention (NCII, 2019, para 2) expands upon the importance of reading screening in schools. This is accomplished through the provision of educational resources on how to adequately screen in schools, and through the compilation of technical information on available reading screening assessments. Most importantly, screening assessments allow schools to efficiently identify students in need of additional academic support, especially when diagnostic accuracy is high. Finally, screening data also assists schools in evaluating school-wide improvement initiatives and aids teachers in making instructional decisions (Albers and Kettler, 2014).

Fuchs, Fuchs, and Compton (2010) suggest screening should be conducted a minimum of two times per year in middle schools that adhere to an RtI framework. Currently, middle school educators report value in screening academic skills. A survey of middle school principals in the Midwest reported 75% of middle schools had practices in place to implement screening at least twice in a year (Prewett, Mellard, Deshler, Allen, Alexander, & Stern, 2012). However, research suggests that there is significant variability in the assessments used by middle schools for academic screening (Brundage, Beckmann-Bartlett & Burns, 2010; Burns & Young, 2018; Stevenson, Reed, & Tighe, 2016). This is made possible by the large number of reading screening assessments available for use at the middle school level. The NCII lists over fifteen of these assessments that may be used by educators (NCII, 2019, academic screening tools chart).

Best Practices in Academic Screening. Albers and Ketler (2014) summarize important considerations for selecting academic screeners. First, schools should determine the purpose of the screening and what academic domain is the focus. Then, it is beneficial to consider what staff qualifications and setting requirements are needed to administer and score the assessment. Schools should review information linking the population for which the assessment is to be used with to the research surrounding the population for which the assessment was created. Most importantly, technical information including reliability, validity and diagnostic accuracy must be reviewed to determine the usefulness of the screening results. These considerations generally fall into two categories: 1) feasibility factors (ease of use, interpretation, and population match) and 2) measurement properties (reliability, validity, and diagnostic accuracy). This

section provides a brief review of feasibility factors; however, the primary focus will be on measurement properties, which is the focus of this thesis.

Feasibility Consideration in Middle School. The National Center on Intensive Instruction (2019) provides summaries on the options schools now have when selecting reading screening assessments. Assessments may be delivered in group settings or individually by student, have online or paper administration formats, timed/untimed assessments, administration times ranging from short one-minute probes up to lengthier half hour assessment, as well as online or in person training opportunities (Christ & Nelson, 2014; National Center on Intensive Intervention, 2019). These are critical considerations as middle schools have limits on resources due to the large number of students being served in each grade. As a result, it may not be feasible to use individualized screeners when group administered assessments result in similar diagnostic accuracy outcomes (Fuchs et al., 2010). Additionally, information provided by the NCII (2019, academic screening tools chart) suggests some screeners require interpretation and scoring by staff trained in specific areas of content or specialties. This may be problematic for middle schools where teachers have certifications in subject areas outside of the focus of the academic screening assessment. As a result, it may be more beneficial for middle schools to consider the use of assessments that are administered and scored online/electronically.

Measurement Properties. Albers and Kettler (2014) as well other authors have highlighted key measurement considerations when evaluating and selecting screening assessments. These authors discuss the importance of using reliable screeners that consistently assess the same skills/content over the course of multiple screenings in an

academic school year. Further, correlations to broader measures demonstrate content validity, or how well the screener's content is in fact measuring the targeted academic domain (Albers & Kettler, 2014; Fuchs, 2004). This is important as higher correlations increase the predictive ability of the screening assessment. Research on these two areas of screening assessment properties is extensive as demonstrated in current research studies and available meta-analyses (Ball & O'Conner, 2016; Kilgus, Methe, Maggin, & Tomasula, 2014; Shin & McMaster, 2019; Yeo, 2010). The results of these research studies and meta-analyses suggest screening assessments in reading not only maintain strong correlations with each other, but also with broader measures such as norm-referenced achievement tests, state tests, and comprehensive reading assessments. However, when screener results are used to make decisions about which specific students are at risk for poor academic outcomes, diagnostic accuracy is one of the most important screener qualities to evaluate. Adequate diagnostic accuracy is important, because schools use screening data to make decisions about how instructional and intervention resources are allocated (Nelson, Norman, & VanDerHeyden, 2017). On the other end of the spectrum, poor screening accuracy may lead to wasted intervention and staffing resources on students that were not truly "at-risk", and/or students continuing to fall behind academically if they are not correctly identified as "at-risk" for future academic challenges (Albers & Kettler, 2014).

The diagnostic accuracy of a screening assessment is dependent on four potential outcomes based on the risk classification. Students may be accurately identified as "not at-risk" (true negatives) or as "at-risk" (true positives) for future academic difficulty. However, assessment can also erroneously classify students as "not at-risk" when they

actually do show later academic difficulties (a false negative) or screeners can erroneously classify students as "at-risk" when they show no future difficulties (a false positive (Albers & Kettler, 2014). Formal evaluations of a screener's diagnostic accuracy typically evaluate the assessment's ability to predict a future outcome, such as performance on a state test. By determining the total number of true negatives, true positives, false negatives, and false positives; indices of sensitivity, specificity, positive predictive value, and negative predictive value can be calculated (Christ & Nelson, 2014; Klingbeil et al., 2015). Sensitivity and specificity are the accurate predictions of students who ended up performing in the "at-risk" (sensitivity) and "not at-risk" (specificity) ranges on an outcome assessment, whereas positive predictive power and negative predictive power evaluates how well the screener determination actually fits the outcome. Specifically, positive predictive power describes the percentage of those who are identified as "at-risk" by a screener and then end up performing in the "not proficient" range on the assessment, and negative predictive power describes the percentage of those who are identified as "not at-risk" by a screener and end up performing in the "proficient" range (Shapiro et al., 2006; Klingbeil et al., 2015; VanDerHeyden, 2011).

Appendix A lists formulas for calculating the above four diagnostic statistics. Diagnostic accuracy indices can range from .00 to 1.00, where a value of .50 indicates a fifty percent accuracy in the risk classification (Klingbeil et al., 2015). Klingbeil et al. (2015) suggests sensitivity should remain above .90 and specificity values ranging from .70 to .80 are acceptable when examining an academic screener's accuracy. Accordingly, accuracy in identifying students who may be "at-risk" for not meeting proficiency standards is emphasized over accuracy in determining students who are "not at-risk".

Finally, neither Compton et al. (2006) nor Klingbeil et al. (2015) provided any threshold suggestions for positive predictive value or negative predictive value as they emphasize the importance of sensitivity and specificity. However, they do suggest predictive values are most useful when the percentage of truly "at-risk" students in the screened sample are similar to the percentage of truly "at-risk" students in the larger student population.

Values for a screening assessment's sensitivity and specificity can also be considered together through area under the curve (AUC) statistical analyses. The NCII (2019, academic screening tools chart) suggests AUC analysis is important and useful for comparing the overall ability of different academic screeners to separate truly "at-risk" students from those who are "not at-risk". Further, information obtained from AUC analysis may help educators determine the best cut-off scores, or levels of performance on the screener, that maximize the screener's accuracy in identifying risk status for students (Klingbeil et al., 2014). AUC values may range from .5 (weak classification) to a 1 (perfect classification). AUC scores are valuable to educators as they indicate how accurate a screening measurement may be when calibrated correctly. This is beneficial information to have when comparing two or more screening assessments especially in situations where resources may only allow for the purchase/delivery of one assessment. Further, schools may benefit by evaluating the AUC of purchased screening assessments locally in order to determine which tool best predicts the academic outcomes of their students.

Middle School Reading Screening Assessments. The majority of academic screeners classified as adequate and appropriate for use in middle schools (NCII, 2019, academic screening tools chart) fall under three categories; fixed achievement tests,

Curriculum Based Measurement (CBM), and Computer Adaptive Tests (CAT). Fixed achievement tests are typically standardized, norm-referenced, academic assessments that have a set number of items a student may complete during testing. The content on these tests have typically been researched to ensure that content validity remains strong with other measures (NCII, 2019, academic screening tools chart). CBM falls under the system of a general outcome measure. General outcome measures are based on the creation of comparable alternate forms from long term objectives (Tindal, 2013). However, the most important feature of CBM when used for screening is that they are brief academic tasks that sample an academic skill. For example, Oral Reading Fluency is a CBM that takes one minute to administer, samples reading automaticity, and is correlated to broader reading measures (Klingbeil et al., 2015; Pearson, 2012). CATs are computerized assessments that present students with a variety of questions related to the academic domain of interest. The adaptive nature of CATs refers to the computer-generated change in difficulty of questions presented based on real time student performance. The overall question bank is typically vetted by a team in order to align with content area educational standards. Current independent research surrounding screening assessments has primarily been completed in elementary settings (Kilgus et al., 2014; Shin, 2019). However, over the past decade these assessments have seen an increase in use across middle schools (Brundage et al., 2010; NCII, 2019, academic screening tools chart)

Middle schools have traditionally used the same assessments as elementary schools when screening academic skills (i.e., curriculum-based measurements and computer adaptive tests), but also include other formats for screening reading

comprehension such as Lexile measures (students read a passage and then are asked to complete questions about the text they read) (Brundage et al., 2010; NCII, 2019, academic screening tools chart; Prewett et al., 2012). However, the recent growth in CATs is particularly meaningful for middle school screening. CATs contain a number of strengths that increase the feasibility of administration at the middle school level. Specifically, CATs are able to be presented in large groups with minimal adult supports. The adaptive nature of CATs increases the range of item difficulty that students may encounter based on their responses. Further CATs are delivered and scored online, which decreases the time a teacher is required to donate for administration and doesn't require school staff to be experts in scoring results (January & Ardoin, 2015). The next sections review the two most commonly administered formats of reading screening assessments in middle schools, CBMs and CATs (NCII, 2019, academic screening tools chart).

Curriculum Based Measurement. Curriculum Based Measurement (CBM) was originally designed to assess academic progress in response to intervention (Deno and Mirkin, 1977). Research supporting CBMs use for academic screening began to accumulate in the 1980s (Tindal, 2013). Reading automaticity, or Oral Reading Fluency (ORF), is one type of CBM that has been heavily researched for use as an academic screener for reading (Espin, 1996, 2007; Fuchs, 2004; Twyman & Tindall, 2007; Wayman, Wallace, Wiley, Ticha & Espin, 2007; Yeo, 2010; Nese, Park, Alonzo, & Tindal, 2011). ORF assesses the automaticity and accuracy a student has when orally reading a short passage (Espin, 1996). Overtime the acceptance of ORF as a valid screening measure has grown and now high percentages of educators endorse ORF as a universal screening assessment for reading (Rowe, Winter, Cook, & Dacruz, 2014).

Further, Klingbeil et al. (2015) examined the diagnostic accuracy of ORF in predicting end of the year reading proficiency on the Measures of Academic Progress for a sample of 548 elementary students. Classification statistics were strong and AUC scores were in the good range, which is consistent with reviews summarized by the NCII (2019, academic screening tools chart).

Less research exists on the use of ORF as a reading screener in middle school. Existing research has primarily focused on psychometric properties and correlations between CBMs and summative assessments. Middle school research suggests R-CBM maintains positive psychometric characteristics in the domains of reliability, content validity, and adequate correlations with well-established reading comprehension assessments; such as, standardized achievement tests and state tests. (Twyman & Tindall, 2007; Ticha, Espin & Wayman, 2009). Further, in a recent meta-analysis, Shin and McMaster (2019) examined this correlational relationship between CBMs (ORF and a reading comprehension measure, MAZE) and student performance on state-wide assessments. The analysis included 61 studies across grades 1 to 10. Results indicated that of all CBMs in reading, ORF provided the highest level of correlations with student performance on state exams, especially at higher grades. However, the strength of correlations between CBM and state testing results decreased overall for secondary students.

There is less research examining the diagnostic accuracy of ORF in middle schools when compared to the body of research that has focused on elementary schools. However, the research that does exist suggests it is best to interpret accuracy through the results of AUC analysis. The NCII (2019, academic tools screening chart)'s review of

existing research and test publisher information pertaining to AUC analysis lists AUC scores as slightly lower for middle school cohorts (compared to elementary cohorts). Specifically, AUC scores were found to generally fall in the "fair" range when predicting future reading performance on state assessments. However, an independent study involving 2943 middle school students reported ORF AUC scores similar to those found in elementary school studies (Baker, Baker, Biancarosa, Park, Bousselot, Smith, & Tindal, 2015). The research presented in this section suggests ORF remains a reliable and valid measure when used on middle school populations. Additionally, evidence suggests ORF may function as a fair reading screener when calibrated correctly. However, it is unclear if ORF will remain the best predictor of performance on state assessments when compared to other formats of academic screeners, such as CATs (Klingbeil et al., 2015).

Computer Adaptive Tests. Computer Adaptive Tests (CATs) are a more recent method of academic screening. The NCII (2019, academic screening tools chart) lists reviews for three different CATs in the area of reading screening; Measures of Academic Progress (MAP), STAR Reading, and FastBridge aReading. Information from test publisher technical manuals indicate CATs have a high level of reliability, provide a valid measure of reading skills that correspond to state standards, and correlate to broader assessments of reading at the middle school level (NCII, 2019, academic screening tools chart); NWEA, 2011). CATs are computer administered assessments that employ Item Response Theory (IRT) as the means to select test items and estimate overall student ability when screening a student's academic skills (Lu and Cong, 2016). Items are typically presented as multiple choice or short answer questions after a student reads a sentence/short paragraph. MAP, of the three CATs reported by the National Center on

Intensive Intervention (NCII, 2019, academic screening tools chart), remains one of the more researched and prolific application of computer adaptive tests for the purpose of reading screening. MAP assess reading comprehension, vocabulary and word structure, and phonics skills (NWEA, 2011). An in-depth description of MAP when used for screening is presented in the methods section.

Area Under the Curve research summarized on the National Center on Intensive Intervention website (NCII, 2019, academic screening tools chart) suggests most CATs, like CBM, have the potential to accurately classify students as either "at-risk" or "not at-risk" in middle school when used as screeners in the area of reading. Further, when test publisher recommended cut-scores are used the CATs listed by the NCII (2019, academic screening tools chart) are slightly better at correctly identifying students who are "not at-risk" versus those that are "at-risk". Specifically, AUC scores, the ability to separate "at-risk" readers from "not at-risk" readers, for middle school samples ranged from .88 to .93; sensitivity from .75 to .82; and specificity from .81 to .88. By applying the presented suggested descriptors from Compton et al. (2006) the reported sensitivity, and specificity of CAT assessments range from fair to good when test publisher suggested cut-scores are used to identify "at-risk" students while AUC range from good to excellent.

Combined Screener Models. Ball and O'Connor (2016) found that diagnostic accuracy (for predicting "at-risk" students in elementary school) increased to the excellent range when CAT reading screener results were paired with CBM reading results in a combined screening model. This study was completed at the elementary level and included 399 second grade students in the state of Wisconsin. The authors suggested that the combination of screening results benefits students needing early intervention, and

assists schools in making larger decisions about how to structure academic supports in schools. As suggested above, accurate screening assessments increase the likelihood that staffing resources and interventions are appropriately allocated in schools by limiting the number of students who are incorrectly identified as needing intervention services. Additionally, correct identification of at-risk students allows schools to intervene early rather than attempt to remediate skills later through intensive services (NCII, 2019, para 2). Therefore, some schools attempt to administer multiple screening assessments in an effort to increase the accuracy in identifying students who are "at-risk" for future reading difficulties (Stevenson, 2017). These procedures have led to new research examining the predictability/use of combined screening models (Nelson et al., 2016). Further research on combined screening models in middle schools is beneficial to educators as it helps to address the need for balance accuracy and feasibility of the screening practices (Fuchs et al., 2010; Nelson et al., 2016).

Klingbiel et al. (2015) examined the predictive validity and diagnostic accuracy of combining Oral Reading Fluency Probes (ORF), Measures of Academic Progress (MAP), and the Fountas and Pinnell Benchmark Assessment System (BAS) screening assessments. Their study involved 548 second and third grade students from four separate elementary schools in the Midwest. Fall screening scores from the three assessments were used to predict student performance on the spring MAP assessment. First, consistent with previous research ORF and MAP scores both showed adequate diagnostic accuracy when considered in isolation. The BAS did not meet the standards suggested in Compton et al. (2016), but, the combination of all three measures led to the best decision making by accounting for 65% of the variance in spring MAP assessment. However, this was

only slightly better than the combination of MAP and ORF, which accounted for 60% of the variance in spring MAP scores. Based on these findings the authors concluded that a combined screening model has the potential to increase the accuracy of screening predictions, but they questioned if the improvement in accuracy of a combined model was worth the additional staffing resources that are needed to administer a second screening assessment to students.

In a similar study in a middle school setting, Nelson et al. (2016) examined the diagnostic accuracy of single and combined screening profiles to predict future state test performance at the middle school level. However, instead of using traditional screening assessments such as CATs or CBMs, the authors created a screening profile off of past state test results and teacher rating scales. Past state test scores were divided into dichotomous categories of "at-risk" / "not at-risk" performances. Cut-scores for determining these two categories were based on either state test publisher recommendations or by local recommendation (local recommendations generally suggested a higher cut-score). Rating scales asked classroom teachers to rate if students had met/ had not met classroom academic standards in reading. Their study included historical information for 641 middle school students, grades six through eight. Predictions based on dichotomous risk categories (state tests), when determined by locally recommended cut-scores, were the best single predictor of future state test performance. Additionally, this single screening model held higher diagnostic accuracy than any combination of combined screening model in the study, Further, the identified AUC values corresponding to screening with past state test scores were equal to those presented earlier for both CBM and CATs. Based on these findings the authors suggested

past state test results may save middle schools financial resources and staffing resources if used for screening in place of more traditional screening assessments. Based on this research it is important to further explore the accuracy of CBM and CATs when combined in order to make a comparison to other methods of combined academic screening.

Summary of Research

The information reviewed in this chapter suggests CBM and CATs are capable screening assessments in the area of reading. CBM currently has the largest body of research to support its use; however, it is unclear if CBM remains the most accurate and feasible screening assessment in middle school. Research on screening with CATs is less prolific but what is available suggests CATs may maintain good levels of diagnostic accuracy at the middle school level when calibrated correctly, while remaining a feasible option. Further, research comparing CBM and CATs on a middle school sample is minimal at this time. More recent research has focused on combined screening models (Klingbiel et al., 2015; Nelson et al., 2015; Stevenson, 2017. However, research examining a combined model of CBM and CAT at the middle school level was not found during the review of existing literature.

Statement of the Problem

Being able to accurately predict a student's future performance on state assessments provides schools with an opportunity to make intelligent decisions about staffing, intervention allocation, and curriculum adjustments. The purpose of this study is to examine the diagnostic accuracy of the Measures of Academic Progress (NWEA, 2011) and AIMSweb Reading-Curriculum Based Measurement (Pearson, 2012) as

individual and combined reading screening assessments for identifying students who are "at-risk" to score in the "not proficient" range on the Minnesota Comprehensive Assessment – Third Edition (Minnesota Department of Education, 2013). Filling this gap in research may provide valuable information to educators contemplating the value of administering multiple or single screening assessments at the middle school level.

The two research questions addressed are:

1) Which screener, R-CBM or MAP, is the best reading screener for students who eventually perform in the "not proficient" range on the Minnesota Comprehensive Assessment Series-III (MCA-III) in 8th grade, as measured by fall screening results?
2) Is diagnostic accuracy improved when a combined screening model is used to predict "not proficient" performances on the MCA-III?

Methods

This chapter describes the methods of the study. First, participant demographics and a description of the educational setting where academic screening took place are described. Next, technical information is presented on both academic screening assessments and the state assessment that are analyzed in this study. Finally, the last two sections are devoted to the procedures and data analysis format used in the current study.

Participants and Setting

School screening data and MCA-III scores were obtained for 8th grade students from four school districts that were part of a larger educational consortium in the Upper Midwest. The 8th grade classes from each district ranged in size from 45 students to 238 students. The districts were part of a larger consortium that assisted with the coordination of specialized services for students, including special education, school psychology, school social work, physical therapy, occupational therapy, and audiology. Three out of the four districts were in a rural setting and the fourth district was in a suburban setting. The suburban district contained the largest middle school serving grades 6-8 and it was 59% of the participant population. The remaining three rural districts had combined middle/high schools that served students in grades 7-12.

A total of 404, 8th grade students, were enrolled in the four districts at the time of the spring MCA-III state assessment. Of those students, 389 had scores on both the MAP and R-CBM reading screeners; data was missing for 15 students, 4% of the initial sample. When reviewed, the demographics of the missing cases did not differ significantly from the original sample. As a result, missing data cases were removed from the test sample without a reduction in power or limits to generalization of results. On average, 93% (SD

= 5.20%) of the overall sample identified as White; 3% (SD = 8.06%) as American Indian; 1% (SD = .96%) as Asian/Pacific Islander; 2% (SD = 1.15%) as Black; and 1% (SD = 1.91%) as Hispanic. Approximately half of the sample was female (M = 49%, SD = 9.02%). Additionally, 5% (SD = 2.00%) received special education services and 37% (SD = 10.90) qualified for free or reduced lunch. No further demographic information was provided by the consortium in relation to the student sample. Demographics by middle school are summarized in Table 1.

Table1

Sample Characteristics

Characteristic	School 1	School 2	School 3	School 4	Total
N	46	238	60	45	389
White/Nonwhite	98% / 2%	96% / 4%	99% / 1%	87% / 13%	93% / 7%
Special Education	<%1	5%	<1%	1%	5%
Female/Male	35% / 65%	49% / 51%	55% / 45%	53% / 47%	49% / 51%
FRL	48%	30%	43%	56%	37%

Note: FRL = Free and Reduced Lunch

Measures

AIMSweb Reading-Curriculum Based Measurement (R-CBM). R-CBM is a 1-minute assessment of oral reading fluency. It assesses reading accuracy and automaticity (Pearson, 2012). Three short passages written at an 8th grade level are administered when used for academic screening in 8th grade. For each administered probe the number of correct words read by the student and errors are recorded.

The student's median words read correctly score on the three probes is used as the overall score. The R-CBM manual suggests 8th grade students performing at or above the 45th percentile rank on national norms are at benchmark and have a high chance of scoring in the proficient range on state tests. Thus, 8th grade students' performing below a raw score of 142 words read correct per minute (R-CBM) in the fall of 2012 were deemed to have failed the screening and were predicted to perform below proficiency on the spring state assessment.

The AIMSweb technical manual (Pearson, 2012) lists 8th grade fall Alternate Form Reliability as .94, interrater reliability as .96 and Standard Error of Measurement (SEM) of R-CBM scores as 6.9, which suggests R-CBM is a highly reliable measure. The mean student performance in the fall of 8th grade is indicated as a raw score of 142 (SD = 34). Additionally, Pearson reports Adjusted Criterion Validity between 8th grade fall R-CBM screening scores and state reading test results as r = .60.

Measures of Academic Progress-Reading (MAP). The Measures of Academic Progress in Reading (MAP; Northwest Evaluation Association, 2011) is an untimed computer adaptive assessment aligned with the Common Core State Standards. Items on MAP Reading assess reading comprehension, vocabulary and word structure, and phonics skills (NWEA, 2011). MAP is administered in a group or individualized setting. Administration time averages 40 minutes per student or group (Center on Response to Intervention, 2019). During screening, multiple-choice questions are presented sequentially. Student performance is scored in real time by the computer software until a student achieves a 1 to 1 ratio of correct to incorrect responses on items that are similar in difficulty (Northwest Evaluation Association, 2011).

Following administration, student performance is converted into Rasch Unit (RIT) scores. A RIT score is a standard achievement score. The mean performance for 8th grade students is identified as a RIT of 219 (SD = 16). Additionally, a RIT of 219 serves as the suggested cut off-score for determining at-risk status. Specifically, the MAP technical manual suggests students performing below the 50th percentile rank (RIT 219) in 8th grade are "at-risk "for performing in the "not proficient" range on the aligned state reading assessment.

The Center on Response to intervention (2019) reports fall to spring MAP Reading test-retest reliability coefficients ranging from .65-.83 for and internal consistency ranging from .76-.82. Additionally, concurrent validity with state accountability tests coefficients range from r = .58-.83 and predicative validity ranges from r = .63-.82.

The Minnesota Comprehensive Assessment Series-III (MCA-III). The Minnesota Comprehensive Assessment Series – III (MCA-III; Minnesota Department of Education, 2018) is a comprehensive state test that assesses student proficiency with academic skills and knowledge identified in the Minnesota Academic Standards. MCA-III is administered annually in the spring to students in grades 3-10 and assesses reading, mathematics, and science. The reading portion of the MCA-III includes multiple-choice questions that assess a student's ability to comprehend and analyze literature and informational texts. Questions may target explicit recall of facts or information, basic synthesizing of text or inferencing, reasoning skills requiring abstract thinking, and sequencing of information (Minnesota Department of Education, 2018).

At the time of the study, the MCA-III was administered online in either a group or individualized setting. All 8th grade students completed the same test items. MCA-III scores included proficiency categories of Does Not Meet the Standards, Partially Meets the Standards, Meets the Standards, and Exceeds the Standards. However, for this study MCA-III scaled scores were converted into a dichotomous "proficient" (Meets the Standards and Exceeds the Standards) and "not proficient" (Does Not Meet the Standards and Partially Meets the Standards) categories. Specifically, scaled scores ranging from 801-849 were designated as "not proficient" and scaled scores ranging from 850-899 were designated as "proficient". Statewide assessment results indicated 54% of 8th grade students in Minnesota who were assessed during the academic year of the study passed the MCA-III in reading (Minnesota Department of Education, 2013).

Procedure

According to district policies, all fall academic screeners were administered between September and October. School staff administered the MAP assessment in a large group setting on individual computers. Students who were absent on the day of screening completed the test individually when they returned to school. Additionally, the school districts used R-CBM probes to screen students reading ability each year during the fall, winter, and spring. During the fall, students completed three probes with a trained staff member in a 1:1 setting. Probes were hand scored and the median value of words read correct in one minute for each student was reported.

The MCA-III is completed in May of each school year. Trained school staff administered the MCA-III according to state guidelines in group or individualized settings. Students completed the assessment on separate computers. Results were scored

electronically before being provided to the district by the Minnesota Department of Education following administration and completion of the assessments. All students took the MCA-III unless they were opted out by a parent or not included because of severe special education impairments.

IRB approval from the University of Wisconsin Eau Claire was obtained prior to the release of data from the consortium. Data was obtained archivally from the consortium serving the four school districts. All identifying information for students was removed prior to being shared. To maintain data security, the data files were encrypted before they were shared. Fidelity of administration for both measures was not included in the archival data from the consortium.

Data Analysis

Descriptive Statistics. Descriptive statistics analyzing the mean, standard deviation, skewness, and kurtosis were completed to analyze score distribution for all measures. Additionally, correlation analyses were completed between MAP and R-CBM, MAP and MCA-III, and R-CBM and MCA-III in order to compare the similarity of scores.

Dichotomous Risk Variables. Dichotomous risk variables for screening assessments were created at the onset of data analysis. MAP and R-CBM cut-off scores for determining risk were based on publishers' research linking their respective screening tools performance to state assessments (Pearson, 2012; Northwest Evaluation Association, 2011). Specifically, fall MAP scores below the 50th percentile (8th grade) and R-CBM scores below the 45th percentile (8th grade) were considered "at-risk". Further, scores below both cut-off scores were required to be identified as "at-risk" on the

combined screening model. These cut-off scores were used in both the classification accuracy and receiver operating characteristic curve analyses detailed below.

Diagnostic Accuracy. Hand calculations of classification accuracy statistics were used to assess the sensitivity, specificity, positive predictive power, and negative predictive power of MAP and R-CBM as individual screeners and a combined screening model (formulas are listed in appendix A). For the combined screening model, students that performed in the "at-risk" range on both MAP and R-CBM were determined to be "at-risk" in the combined model. As a result, students who scored below the 50th percentile on MAP and below the 45th percentile on R-CBM were deemed "at-risk" in the combined model.

Area Under the Curve. Receiver operating characteristic (ROC) curves were used to inspect the diagnostic accuracy of each screening assessment individually as well as when combined. Analyses were conducted using SPSS V24.0 (IBM Corp., 2016). The ROC analysis includes measures of sensitivity and specificity to determine the area under the curve (AUC). AUC is used in classification analysis to compare the accuracy of predictions of differing models (Compton et al., 2006; Klingbeil et al., 2015). In this study, the ROC curves plotted the rates of correctly identifying a "not proficient" performance (true positives) on the MCA-III against incorrectly identifying a "not proficient" performance (false positives). AUC statistics can range from .5, poor diagnostic capability, to 1, perfect diagnostic capability, (Compton et al., 2006; Klingbeil et al., 2015). Compton et al. (2006) suggests ROC curve values >.90 are excellent, values ranging from .80 to .90 are good, .70 to .80 are fair, and below .70 are poor when assessing a tool's diagnostic ability.

Results

Chapter three describes results of conducted analyses. Analyses consisted of descriptive statistics, correlations, classification accuracy calculations, and receiver operating characteristic (ROC) curve. Statistical results and corresponding tables of data are presented below.

Descriptive Statistics and Correlations

Descriptive statistics for all three measures are shown in Table 2 and Table 3. No alterations to R-CBM or MAP scores were required because students with missing data were excluded. The mean performance in the fall for 8th grade students on R-CBM and MAP was respectively 163 words read correct per minute (range = 60 - 281) and RIT 223 (range = 184 - 255). For the fall screening, 68% of students performed in the proficient range on MAP and 76% performed in the proficient range on R-CBM. In the Spring approximately 51% of students scored in the proficient range on the MCA-III assessment in reading. The mean MCA-III performance was a scaled score of 850 with scores ranging from 802-898. Visual analysis of scatter plots demonstrated no outlying scores and all three variables produced normal distributions of scores with minimal skew when assessed for kurtosis and skewness. Finally, moderate correlations were observed between MAP and R-CBM ($r = .62$) and R-CBM and MCA ($r = .55$), and moderate to high correlation for MAP and the MCA-III ($r = .73$; Table 3). These values fall within the ranges of screener to state test correlations as reported by the test publishers (NWEA, 2011; Pearson, 2012).

Table 2

Descriptive Statistics for Reading Screeners and State Test (N = 389)

	M	*SD*	Skewness	Kurtosis	Not Proficient
Screener					
Fall R-CBM	163.00	32.15	.11	.73	24.00%
Fall MAP	223.00	11.66	-.36	.27	32.00%
State Assessment					
Spr MCA-III	850.00	16.21	-.20	.73	49.00%

Note. R-CBM = Reading Curriculum Based Measurement (AIMSweb); MAP = Measures of Academic Progress Reading (NWEA); MCA-III = Minnesota Comprehensive Assessment-Series III in Reading.

Table 3

Correlation Between Variables

	r
R-CBM and MAP	.62
R-CBM and MCA-III	.55
MAP and MCA-III	.73

Research Question 1: Single Screener Models

The first research question examined the diagnostic accuracy of each individual screener, MAP and R-CBM, in predicting spring MCA-III performance when using publisher recommended cut scores (Table 4). MAP screening resulted in excellent specificity of .94; however, the resulting sensitivity value of .59 did not reach the minimum recommended .80. R-CBM screening also resulted in excellent specificity of .92 but low sensitivity .41. Poor sensitivity values indicate many students were

misclassified as "not at-risk" during screening, but subsequently performed below proficiency on the MCA-III. Further, both measures demonstrated high positive predictive values (low numbers of false positives) and low negative predictive values (high rates of false negatives).

When compared, MAP was superior to R-CBM in the categories of specificity and sensitivity. Additionally, MAP scores resulted in good diagnostic accuracy (AUC .90), while R-CBM resulted in fair diagnostic accuracy (AUC .79). AUC classifications categories were provided by (Compton et al., 2006). Values may be classified as follows: >.90 are excellent, values ranging from .80 to .90 are good, .70 to .80 are fair, and below .70 are poor.

Table 4

Diagnostic Accuracy of MAP and R-CBM Universal Screening Measures (N=389)

	TP	TN	FP	FN	AUC	Sensitivity	Specificity	PPV	NPV
Fall MAP	112	188	11	78	.90	.59	.94	.91	.71
Fall R-CBM	78	184	15	112	.79	.41	.92	.84	.62
MAP+R-CBM	63	197	2	127	.84	.33	.99	.97	.61

Note. MAP=Measures of Academic Progress; R-CBM-Reading Curriculum Based Measurement; MAP+R-CBM = Combination of Measures of Academic Progress and Reading Curriculum Based Measurement; TP = true positives; TN = true negatives; FP = false positives; FN = false negatives; AUC = area under the curve; sensitivity = TP / (TP + FN); specificity = TN / (TN + FP); PPV = positive predictive value; PPV = TP / (TP + FP); NPV = negative predictive value; NPV = TN / (TN + FN).

Research Question 2: Screener Combination

The second research question examined the diagnostic accuracy of using the combination of both assessments to screen for future performance on the MCA-III. The

combination of screeners focused on identifying students as "at-risk" if a student obtained an "at-risk" score on both screeners. The combined diagnostic accuracy of these screeners was in the good range with an AUC value of .84, but below the overall accuracy when MAP was used as a single screener model. Specificity (.99) and PPV (.97) remained excellent; however, sensitivity (.33) and NPV (.61) were poor. The combination of MAP and R-CBM screening results led to an increase (.05) in specificity and (.06) in PPV, when compared to MAP as a single screening assessment. However, a decrease (.26) in sensitivity and (.10) in NPV were observed with the two-measure model.

Discussion

This final chapter discusses the study's findings and implications. Specifically, implications for reading screening in middle school, screening with MAP and R-CBM assessments, and screening with a combination of assessments and data are explored. Finally, limitations of the study and recommendations for potential future research are reviewed.

Universal academic screening is a focus of middle schools implementing RtI, and it is used by educators making data-based decisions about resource allocation to support students (Brundage, Beckmann-Bartlett, & Burns, 2010; Fuchs et al., 2010; Prewett, Mellard, Deshler, Allen, Alexander, & Stern, 2012). This study compared the diagnostic accuracy of two single screening measures and one combined screening model of reading for 8th grade students. The focus was to determine which model most accurately predicted students who were "at-risk" to perform in the "not proficient" range on the end of the year MCA-III state test.

Interpretations

Research Question 1: Single Screener Models. The first research question analyzed which individual assessment, MAP or R-CBM, acted as a better diagnostic screener when predicting future performance on the MCA-III. Hand calculations of diagnostic accuracy indicated students' fall MAP and R-CBM performance resulted in excellent levels of specificity. Specifically, they were both accurate in identifying students who were on track to score in the proficient range on the state test. However, both assessments demonstrated poor sensitivity and were inaccurate in identifying students who were truly at-risk to perform in the "not proficient" range on the state test.

These results suggest that both screeners failed to identify many students who subsequently performed below expectations on the MCA-III. This is problematic because missing students (i.e., having many false negatives) can lead to students not receiving needed interventions and supports, thus perpetuating their academic weaknesses (Klingbeil et al., 2015).

This is an interesting finding as a more liberal normative cut score was used in the analysis. For example, NCII (2019, academic screening tools chart) suggests schools often use a three-tier classification system for screeners that includes the categories of significant risk, moderate risk, and not at-risk. The selected cut score in this study aligned with the threshold that typically falls between the "some risk" and "not at-risk" categories. (NWEA, 2011; Pearson, 2012). Selecting this cut score, versus the lower cut score that is used in a three-tier model, is typically hypothesized to increase the sensitivity when screening for "at-risk" status (Compton, Fuchs, Fuchs, Bouton, Gilbert, Barguero, Cho, & Crouch, 2010). The increase of students being identified as "at-risk" is due to widening the range of scores on the screener that will result in an "at-risk" identification. However, recent research suggests the sensitivity of screening assessments may be negatively impacted when test publisher recommended cut-scores are used to define "at-risk" thresholds (Kilgus et al., 2014; Nelson et al., 2017). Further, these researchers found that calibrating screening assessment cut-off scores through the use of local data typically leads to an increase in the sensitivity of screening assessments when identifying "at-risk" students. Specifically, Kilgus et al. (2014) found that R-CBM has the potential to yield higher sensitivity and lower false negatives when cut-scores are based on past screening to outcome relationships. In this example, historical data

suggested there was a trend that the publisher recommended cut-score was too low as numerous "at-risk" students were repeatedly being misclassified. The data suggested a higher cut-score may result in the screening assessment having better diagnostic accuracy for the targeted school population. This is true of MAP as well given a review of information provided by the test publisher. According to data from the technical manual, sensitivity values of .58 (below recommended value of .90) and specificity values of .89 (above recommended value of .80) are typical when the publisher recommended cut-off score (50^{th} percentile rank) is used for 8^{th} grade students (NWEA, 2012).

Area under the curve (AUC) results made it possible to examine the diagnostic accuracy of both MAP and R-CBM when a specific cut-off is not designated. AUC analysis plots the sensitivity/specificity pairs according to each potential cut-off score in the predictor variable. The analysis is then able to locate the score that produced the most accurate classification of "at-risk" and "not at-risk" students (Klingbeil et al., 2014). AUC statistics can range from .5 (poor diagnostic capability) to 1 (perfect diagnostic power) where ROC curve values >.90 are excellent, values ranging from .80 to .90 are good, .70 to .80 are fair, and below .70 are poor when assessing a tool's diagnostic ability (Compton et al., 2006; Klingbeil et al., 2015). MAP demonstrated the better performance and the AUC value of .90 suggests good diagnostic accuracy. MAP's high AUC indicates that for this sample of students there is a point where MAP can correctly predict students as "at-risk" / "not-at-risk" with 90% accuracy. In contrast R-CBM's AUC value of .79 suggests it is near the border between fair and good, which means that at best, a different cut score would only accurately predict 79% of the student cases as "at-risk" / "not at-risk". Further, AUC results are consistent with recommendations that locally derived cut-

scores are likely to provide the most accurate diagnostic accuracy when used for screening. This finding is important as increasing accuracy leads to better identification of students who are "at-risk" and may require intervention or skill remediation. Further, accurate screening helps to prevent academic gap growth, which reduces the need for greater resource allocation in the future (Prewett et al., 2012).

The final intent of single screener analysis was to compare the diagnostic accuracy of MAP and R-CBM in 8th grade. As stated above overall AUC scores and hand calculations of diagnostic accuracy suggest MAP is a more accurate screening tool for identifying students "at-risk" for poor reading performance in 8th grade. Previous research shows that R-CBM typically loses diagnostic accuracy when used as a screener with older students, especially students in middle school and beyond (Shin & McMaster, 2019). A similar loss of screening accuracy was found in a meta-analysis completed by Kilgus et al. (2014) when R-CBM was used to predict performance on state tests at the middle school level. As a result, the current findings of low diagnostic accuracy for screening with R-CBM in middle school is not a surprise.

Kilgus et al. (2014) also discussed screener accuracy relative to base rates of the predicted outcome in the general population. That is, they concluded R-CBM is an ideal diagnostic tool when the base rate of students performing in the "not proficient" range is relatively high. True base rates in a sample do not have an impact on the sensitivity or specificity of a tool but can significantly impact the positive and negative predictive value of a screening instrument (Glover & Albers, 2007; Labarge, McCaffery, & Brown, 2001). For example, screenings completed with student populations where most students end up performing in the proficient range on the outcome variable of interest (e.g., state

tests) are poor at identifying students who are "at-risk" without over estimating (false positives) the number of "at-risk" students (Glover and Albers, 2007). The sample analyzed in this study did not have a relatively high base rate of "not proficient" scores (49% scored in the not proficient range on the MCA-III), which may partially explain the lower accuracy of R-CBM. Further, in the current study MAP out performed R-CBM in both positive predictive value and negative predictive value. Positive predictive value being the ability to identify students as "at-risk" who then actually go on to fail the state test, and negative predictive value being the ability to identify students as "not at-risk" who then actually go on to pass the state test. The importance of these values lies in what the emphasis of the screening is. If the purpose of screening is to identify only those who are truly "at-risk" then it is beneficial to select a screening assessment with a high positive predictive value. However, if the goal is to identify only those who are truly "not at-risk" then selecting a tool with a high negative predictive value will be more impactful (Compton et al., 2010). Being that R-CBM underperformed in both values it appears MAP would be the better assessment for either of these two scenarios.

Research Question 2: Combined Screener Model. The second research question analyzed whether combining MAP and R-CBM enhanced overall diagnostic accuracy in predicting success on the MCA-III. This analysis is relevant for academic screening practices as results evaluate the value of administering multiple or single screening assessments at the middle school level. Given that MAP was identified as the more accurate screening tool, it was used as the primary screening assessment in the combined screener model. Students who performed in the "at-risk" range on MAP were then re-screened using R-CBM. When calculating diagnostic accuracy outcomes,

students identified as "at-risk" on both MAP and R-CBM were then predicted to perform in the "not proficient" range on the MCA-III.

For the current sample of middle school students, the combined model increased the specificity of identifying students who later performed poorly on the MCA-III but greatly reduced the sensitivity relative to relying on MAP results alone. In the combined model, false negatives were even higher suggesting that even more students potentially in need of academic support were not identified as "at-risk" when two screeners were used. Finally, the potential diagnostic accuracy of the combined screening model as measured by ROC analysis was also lower in comparison to MAP alone, but nearly equal to R-CBM. In summary, the combined screening model increased the ability to locate students who were truly "not at-risk" for failing the state test and the overall predicative power of the combined assessments decreased when compared to the results of MAP alone.

However, some areas of accuracy improved in the combine model. The combined model lead to a positive predictive value that was greater than what was produced in the single screening models. In this instance the second screening assessment acted as a confirmation that the initial screening result was accurate in identifying only those who were truly "at-risk" for failing the future state test. This outcome best benefits schools that need to be certain that interventions are only being delivered to students who truly need them due to limited staff and intensive intervention resources. Including R-CBM only led to nine additional students being correctly identified as "not at-risk" from a sample of over 300. This calls into question the practicality of adding R-CBM when the return in added diagnostic accuracy is so small (versus the additional amount of time and resources required to perform a second screening).

The findings of decreased screening accuracy with a combined screening model are not consistent with Klingbeil (2015). Instead, Klingbeil (2015) determined that using multiple screeners increased the overall diagnostic accuracy when attempting to predict "at-risk" status with a sample of elementary school students. It is important to note that MAP, a correlate to state assessments, was used as the outcome measure in Klingbeil (2015). However, the biggest difference between Klingbeil (2015) and this current study is how students were identified as "at-risk" in the respective combined screening models. This current study required students to fail both screeners to be identified as "at-risk". In Klingbeil (2015) students were "at-risk" if they failed the first round of screening or the second round of screening. This difference explains the divergence of findings for the current study and what was found in Klingbeil (2015). Specifically, the more stringent approach of the present study led to a decrease in the number of false positives and an increase in the number of false negatives that were observed in the single screening models. Conversely, the more liberal approach in Klingbeil (2015) had the opposite impact, a decrease in false negatives but an increase in false positives. When compared, the approach taken in Klingbeil (2015) is the more useful of the two models when the goal is to prevent the misclassification of truly "at-risk" students as "not at-risk". Future research could replicate the model used in Klingbeil (2015) on middle school samples in order to verify what was found in the elementary setting.

Implications for Practice and Future Research

Achieving optimal accuracy in a single screener model. The above discussion of results suggests a few overarching implications for RtI systems in a middle school setting. First, analysis of screener sensitivity did not support the use of either MAP or R-

CBM as a single screening tool when attempting to identify students who are at-risk to perform below the proficiency cut-off score on state tests. If the suggested test publisher cut-off score was used as the sole predicator of at-risk status in the schools sampled, a large number of students who need additional support would be missed. This was found to be true in other similar studies (NWEA, 2012; Klingbeil, 2015; Stevenson, 2017). However, AUC analysis suggests MAP has the potential to be a good diagnostic assessment of at-risk status in reading when calibrated correctly. Future middle school research should continue to explore MAP's diagnostic accuracy for cut scores that are calibrated to their student's performance on state tests.

Practicality of a combined screening model. Findings did not support the use of a two-screener model when test publisher recommended normative cut-off scores are used. Once again schools will be missing larger numbers of students who need intervention. Stevenson (2017) also found similar results when using multiple CBM screeners, Further, AUC analysis suggests the potential diagnostic accuracy when combining MAP and R-CBM screenings still did not surpass the potential of MAP alone. As a result, a combined screening model may result in lower diagnostic accuracy, a higher staff resource cost for the administration of additional screeners, and an additional monetary cost for the purchase of a second screening assessment, when combining these two reading assessments at the middle school level. However, it should be noted that this finding only relates to a screening model that requires failing scores on both screening assessments. For example, future middle school research could explore combined screening models that require only one failing score in order for a student to be deemed "at-risk". Further, future research could explore using higher cut-score thresholds

(expanding the range of scores on a screening assessment that deem a student as "at-risk") than what is recommended by test publishers for use in combined screener models.

Inclusion of existing data in at-risk identification. Educators at the middle school level must weigh the costs and benefits of using traditional screening practices commonly implemented in elementary schools. Fuchs et al. (2010) suggests RtI at the middle school level may benefit from deviating from the typical elementary approach. Screening assessments completed multiple times a year may not be necessary because student records contain a host of academic information by the time students enter middle school. For example, Stevenson (2016) demonstrated the usefulness of past state assessment findings and anecdotal teacher observation. Specifically, he found that past state assessment results in combination with qualitative teacher ratings on student reading performance had adequate levels of predictive power when determining at-risk status for performing in the "not proficient range" on future state assessments. Further, in a follow up study, Stevenson (2017), added that Oral Reading Fluency (ORF) and a more detailed reading comprehension CBM did not yield an advantage over existing data (attendance records, grades, discipline referrals, and past state test results) when making classification decisions.

The above findings make sense when consideration is given to how schedules may differ between elementary schools and middles schools. Middle schools often have multiple periods (potentially even six or seven different periods) in a day that are taught by different staff members. This may increase the variability in each individual student's schedule, and so waiting for fall screening data may be too late to alter schedules in order to deliver interventions to students who are flagged as "at-risk". It may be more feasible

to use existing data to set schedules prior to the beginning of the school year in these cases. This will likely reduce the wait time before students can begin intervention while also better identifying the "at-risk" students who are missed when a single screener is used.

Limitations

Findings from the current study should be interpreted with regards to respective limitations. First, information was not available on the fidelity of administration of the screening assessments. Without this information, there no guarantee that screening assessments were completed as intended. As a result, the validity of the scores obtained during screening cannot be guaranteed. A reduction in fidelity would decrease the ability to generalize results to other districts. Future research should aim to have more direct input into the monitoring of screening fidelity.

Second, information was not available to determine whether students received additional reading support during the period following the fall screening and before MCA-III assessments in the spring. However, students in middle school with larger skill deficits often require more time in intervention in comparison to elementary students (Vaughn and Fletcher, 2012) and are thus less likely to change their risk status over the course of a year. Future research would do well to control for intervention status in order to determine the impact a year of intervention has on the accuracy of fall to spring classification accuracy.

A final limitation of this study is due to the lack of information pertaining to the percentage of English Learners in the observed sample. As a result, analyses did not evaluate screening accuracy differences for English Learners. Future research should

examine the diagnostic accuracy of these assessments for use with middle school students with different levels of English language proficiency.

Conclusion

Universal screening continues to be a common practice for identifying students who are at-risk for falling below proficiency on grade level standards. The results of this study demonstrate that MAP was the better individual measure when assessing diagnostic accuracy. Further the combination of screeners did not improve the diagnostic utility in comparison to MAP alone. However, these results suggest using only MAP to screen students in 8th grade may result in many students being misidentified as not at-risk when using publisher recommended cut-off scores. Emerging evidence from other research suggests past data, the modification of cut-off scores, and/or including additional grouping categories for borderline students may increase the accurate identification of risk (Klingbeil et al., 2015; American Institute for Research, 2019). Additional research is warranted to examine other screening approaches in coordination with MAP at the middle school level.

References

Albers, C. A., & Kettler, R., J. (2014). Best practices in universal screening. In Harrison, P., L, & Thomas, A. (Eds.), *Best practices in school psychology: Data-based and collaborative decision making* (pp 121-131). Bethesda, MD: National Association of School Psychologists.

Baker, D., Baker, S., Biancarosa, G., Park, B., Bousselot, T., Smith, J.-L., Tindal, G. (2015). Validity of CBM measures of oral reading fluency and reading comprehension on high-stakes reading assessments in Grades 7 and 8. *Reading & Writing*, *28*(1), 57–104.

Ball, C. R., & O'Connor, E. (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.

Brundage, A., Beckmann-Bartlett, C., & Burns, M. K. (2010) Response to intervention: Alice Birney middle school's model, experience, and results. *Communique (0164775X), 39*(1), 1-11).

Burns, M. K., & Young, H. (2018). Test review: Measures of academic progress skills. *Journal of Psychoeducational Assessment. 00*(0), 1-4.

Calhoon, M. (2008). Curriculum-based measurement for mathematics at the high school level: What we do not know...what we need to know. *Assessment for Effective Intervention*, *33*(4), 234-239.

Center for Response to Intervention. Retrived July 07, 2019 form https://rti4success.org/

Christ, T., & Nelson, P. (2014). Screening assessment: Practical and psychometric considerations. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.), *Universal Screening in educational settings: Evidence-based decision making for schools* (pp. 79-110). Washington, DC: American Psychological Association

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.

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

Conley, D. T. (2007). Redefining college readiness. Eugene, OR: Educational Policy Improvement Center. (3)

Deno, S., Mirkin, P. (1977). Data based program modification: A manual. Minneapolis, MN: Leadership Training Institute for Special Education.

Espin, C., Wallace, T., Lembke, E. , Campbell, H., & Long, J. (2010). Creating a progress-monitoring system in reading for middle-school students: Tracking progress toward meeting high-stakes standards. *Learning Disabilities Research & Practice (Blackwell Publishing Limited)*, *25*(2), 60-75.

Fuchs, L. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review*, *33*(2), 188-192.

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2010). Rethinking response to intervention at middle and high school. *School Psychology Review, 39*(1), 22-28.

Gewertz, C. (2011). College for all reconsidered: Are four-year degrees for all?. *Education Week, 30*(34), 6-8.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45*(2), 117-135.

January, S. A., & Ardoin, S. P. (2015). Technical adequacy and acceptability of curriculum-based measurement and the measures of academic progress. *Assessment for Effective Intervention. 41*(1), 3-15.

Johnson, E. S., & Smith, L. A. (2011) Response to intervention in middle school: A case story. *Middle School Journal, 42*(3), 24-32.

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomsula, 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*(2014), 377-405.

Klingbeil, D., McComas, J., Bruns, M., & Helman, L. (2015). Comparison of predictive validity and diagnostic accuracy of screening measures of reading skills. *Psychology in Schools, 52*(5), 500-514.

Labarge, L. S., McCaffrey, R. J., & Brown, T. A. (2003). Neuropsychologists' abilities to determine the predictive value of diagnostic tests. *Archives of Clinical Neuropsychology, 19*(2), 165-175.

Lowman, J. (2006). Poster 16. School principal's views of no child left behind, the achievement gap, and student groups assessed by nclb. *Conference Papers -- American Sociological Association*, 1-5.

Lu, P., Cong, X. (2016) The research on computerized adaptive testing. *Journal of Physics: Conference Series, 710,* 1-10.

Minnesota Department of Education. (2013). Minnesota interpretive guide for Minnesota Comprehensive Assessment – Third Edition. MN.

National Center on Intensive Intervention. (2019). Intensive intervention & multi-tiered system of supports. Retrieved from https://intensiveintervention.org/intensive-intervention/multi-tiered-systems-support.

Nese, J., Park, B., Alonzo, J., & Tindal, G. (2011). Applied curriculum-based measurement as a predictor of high-stakes assessment. *Elementary School Journal, 111*(4), 608-624.

Northwest Evaluation Association. (2011). Technical manual for Measures of Academic Progress™ and Measures of Academic Progress for primary grades™. Lake Oswego, OR:.

Pearson Inc. (2012). Aimsweb technical manual. Bloomington, MN:.

Prewett, S., Mellard, D. F., Deshler, D. D., Allen, J., Alexander, R., & Stern, A. (2012). Response to intervention in middle schools: Practices and outcomes. *Learning Disabilities Research & Practice (Wiley-Blackwell), 27*(3), 136-147.

Rowe, S. S., Witmer, S., Cook, E., & DaCruz, K. (2014). Teachers' attitudes about using curriculum-based measurement in reading (CBM-R) for universal screening and progress monitoring. *Journal of Applied School Psychology 30*(4) (2014): 305-37.

Sanger, D. , Friedli, C. , Brunken, C. , Snow, P. , & Ritzman, M. (2012). Educators' year long reactions to the implementation of a response to intervention (rti) model. *Journal of Ethnographic & Qualitative Research, 7*(2), 98-107.

Shapiro, E., Keller, M , Lutz, J., Santoro, L., & Hintze, J. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment, 24*(1), 19-35.

Shin, J., McMaster, K. (2019). Relations between CBM (oral reading and maze) and reading comprehension on state achievement tests: A meta-analysis. *Journal of School Psychology, 73*(2019), 131-149

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

Stecker, P. M., Lembke, E. S., & Foegen, A. (2008). Using progress monitoring data to improve instructional decision making. *Preventing School Failure, 52*(2), 48-58.

Stevenson, S. A., Reed, D. K., & Tighe, E. L. (2016). Examining potential bias in screening measures for middle school students by special education and low socioeconomic status subgroups. *Psychology in the Schools, 53*(5), 533-547.

Ticha, R., Espin, C., & Wayman, M. (2009). Reading progress monitoring for secondary-school students: Reliability, validity, and sensitivity to growth of reading-aloud and maze-selection measures. *Learning Disabilities Research & Practice, 24*(3), 132-142.

Tindal, G. (2013). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *ISRN Education 2013,* 1-29.

Twyman, T., & Tindall, G. (2007). Extending curriculum-based measurement into middle/secondary schools: The technical adequacy of the concept maze. *Journal of Applied School Psychology*, *24*(1), 49-67.

U.S. Department of Education (2016). Institute of education sciences, National Center for Education Statistics.

VanDerHeyden, A.M. (2011). Technical adequacy of RtI decisions. *Exceptional Children, 77(3),* 335-350.

Vaughn, S., & Fletcher, J.M. (2012). Response to intervention with secondary school students with reading difficulties. *Journal of Learning Disabilities, 45*(3), 244-256.

Wayman, M., Wallace, T. , Wiley, H. , Ticha, R. , & Espin, C. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education*, *41*(2), 85-120.

Williams, R, Ari, O., & Santamaria, C. (2011). Measuring college students' reading comprehension ability using cloze tests. *Journal of Research in Reading*, *34*(2), 215-231.

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.

Appendix A

Definitions of Key Terms in Relation to Academic Screening

Area under the curve (AUC) is an analysis when a statistical curve is obtained by plotting all sensitivity/specificity pairs according to each potential score from the screening assessment. Once these points are plotted the optimal cut-off score is determined based on how accurately the measurement separates cases that have the targeted condition (in screening research these are typically "at-risk" students) from those that do not at the identified cut-off score (Klingbeil et al., 2014)

Computer adaptive tests (CATs) are computerized assessments that adapt in difficulty based on in real time student responses (Klingbeil et al., 2015).

Curriculum based measurement (CBM) is a type of general outcome measure that quickly and accurately samples a targeted academic skill (Shapiro et al., 2006).

Diagnostic accuracy is the ability of a screening assessment to discriminate between "at-risk" and "not at-risk" readers (Compton et al., 2006).

Negative predictive value (NPV) is the percentage of students who were labeled as "not at-risk" on the screening assessment and then ended up performing in the "proficient" range on the state test (Compton et al., 2006).

Positive predictive value (PPV) is the percentage of those who were labeled "at-risk" on the screening assessment and then ended up performing in the "not proficient" range on the state test (Compton et al., 2006).

Receiver operating characteristic (ROC) curve is a plot of the true- positive rate of detecting "at-risk readers" against the corresponding false-positive rate of misidentifying "not at-risk" readers (Klingbeil et al., 2015).

Sensitivity is a screener's ability to correctly identify students who go on to perform in the "not proficient" range on the state test (Klingbeil et al., 2015).

Specificity is how well the screener correctly identified students who went on to meet or exceed the proficient threshold on the state test (Klingbeil et al., 2015).

Appendix B

Formulas Used in Hand Calculation Analysis.

Sensitivity refers to a screener's ability to correctly identify students who test positive on the outcome measure (e.g. score in the "not proficient" range). Sensitivity = Total Number True Positives / (Total Number True Positives + Total Number False Negatives).

Specificity referred to how well the screener correctly identified students who went on to meet or exceed the proficient threshold on the MCA-III (scored in the "proficient" range). Specificity = Total Number True Negatives / (Total Number True Negatives + Total Number False Positives).

Positive predictive value described the percentage of those who were labeled "at-risk" and then ended up performing in the "not proficient" range on the MCA-III. Positive predictive value = Total Number True Positives / (Total Number True Positives + Total Number False Positives).

Negative predictive value described the percentage of those who were labeled as "not at-risk" and ended up performing in the "proficient" range. Negative predictive value = Total Number True Negatives / (Total Number True Negatives + Total Number False Negatives).