

Examining the Diagnostic Accuracy and Predictive Validity for
FastBridge aReading and CBM-R on High Stakes State Tests

By

Danielle M. Swedeen

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

Educational Specialist
School Psychology

At

The University of Wisconsin-Eau Claire

June, 2022

Graduate Studies

The members of the Committee approve the thesis of

Danielle Swedeen presented on June 20th, 2022

Dr. Mary Beth Tusing

Dr. Melissa Coolong-Chaffin

Dr. Brian Orr

APPROVED: _____
Dean of Graduate Studies

Examining the Diagnostic Accuracy and Predictive Validity for
FastBridge aReading and CBM-R on High Stakes States Tests

By

Danielle Swedeen

The University of Wisconsin-Eau Claire, 2022
Under the Supervision of Dr. Mary Beth Tusing

The current study examined the diagnostic accuracy of two common screening assessments in reading, FastBridge Adaptive Reading (aReading) and Curriculum Based Measurement of Oral Reading (CBM-R) when used to predict future performance on the Minnesota Comprehensive Assessment – Third Edition (MCA-III) in spring of 3rd grade. The sample consisted of 164 students enrolled in a school district in the Upper Midwest. Both screeners were moderately correlated with the MCA and had strong diagnostic accuracy; however, both screeners showed poorer sensitivity when vendor cut scores were applied. These results demonstrate the importance of considering local norms for the most accurate screening decisions. Future research could explore the diagnostic accuracy of both screeners combined, using other data to inform decision making, or exploring different cut-scores for defining “at-risk” students.

Thesis Advisor

Date

TABLE OF CONTENTS

	Page
LIST OF TABLES	V
Chapter	
I. LITERATURE REVIEW	1
MTSS and Universal Screening	2
Best Practices in Academic Screening	3
Curriculum Based Measurement as a Universal Screener	5
Computer Adaptive Tests as a Universal Screener	8
National Center for Intensive Intervention (NCII) Review	9
Statement of the Problem	12
II. METHODS	14
Participants and Setting	14
Measures	15
Procedure	19
Data Analysis	19
III. RESULTS	22
Descriptive Statistics and Correlations	22
Research Question	23
IV. DISCUSSION	25
Implications for Practice	28
Implications for Future Research	30
Limitations	31
Conclusion	33
REFERENCES	34

LIST OF TABLES

Table	Page
1. Sample Characteristics (N=164)	15
2. Descriptive Statistics for Reading Screeners and State Test (N = 164)	22
3. Correlation Between Variables	23
4. Diagnostic Accuracy of aReading and R-CBM Universal Screening measures (N=164)	24

Chapter 1

Literature Review

The Every Student Succeeds Act (ESSA) aims to ensure all students are taught to high academic standards. Annual statewide assessments are one way students' progress toward these high standards are measured (Every Student Succeeds Act, 2015). Given the importance of student performance on state outcome assessments, many school districts have made a shift from being reactive to learning deficits identified by state tests to a more preventative approach. Data-based decision-making processes, more commonly referred to as Response to Intervention (RtI) or Multi-Tiered Systems of Support (MTSS) are now common in most schools (Sanger, Friedli, Brunken, Snow, & Ritzman, 2012). RtI consists of different levels or tiers of support, hence MTSS (Castillo, 2014). Early intervention is an essential component of MTSS and universal, or school-wide, screening is often used by schools to identify students at risk for poor learning outcomes early in their educational career so that additional support can be provided (Center on Multi-Tiered Systems of Supports, 2021).

Many methods of academic screening for data-based decision-making have been developed. Curriculum-based measurement (CBM) is one type of assessment that has been established as a reliable and valid indicator of future academic performance (Deno, 2003; Shapiro & Gebhardt, 2012). Various studies have demonstrated positive correlations between CBM scores and future performance on high-stakes assessments, like statewide tests (Deno, 2003; Yeo, 2010). Another viable option for universal screening is computer-adaptive testing (CAT; Shapiro & Gebhardt, 2012). CATs are assessments that adapt to a student's ability level based on their responses to items.

Although the use of CATs as screening tools is more recent addition to academic screening practices, current research on CATs suggests they are accurate predictors of future performance on high-stakes state tests (Ball & O'Connor, 2016). The reliability and validity, generalizability, and efficiency of screeners are critical factors to consider when being used to identify students who are at risk for future learning difficulties (Albers & Kettler, 2014; Deno, 1985). The current study examines the diagnostic accuracy and predictive validity FastBridge's CBM of oral reading fluency (CBMReading) and CAT reading assessment (aReading).

MTSS and Universal Academic Screening

According to Stoiber (2014), MTSS refers to a multi-component, comprehensive, and cohesive school-wide and classroom-based positive support system through which students at risk for academic and behavioral difficulties are identified and provided with evidence-based and data-informed instruction, support, and instruction. Although MTSS models vary, this framework generally consists of three tiers often called primary or universal (Tier 1), secondary or targeted (Tier 2), and tertiary or intensive (Tier 3). Universal academic screening is one essential component to identify those who are not responding to the core instruction at the Tier 1 level and are at risk for future academic difficulties.

According to Albers and Kettler (2014), academic difficulties are often defined as not showing proficiency on summative, or year-end academic assessments such as high stakes state tests. Identifying at-risk students early opens the door for additional intervention services to reduce or even negate academic difficulties. Academic screening also provides educators with data on an individual's or systems' current state and

academic needs. It allows insight into the effectiveness of instruction at the classroom, grade, school, and district level as well as the effectiveness of early intervention services on predicting future academic difficulties. The practice of universal screening in schools typically consists of administering brief assessments multiple times throughout the academic year (fall, winter, spring) to identify specific students at risk of poor performance on end-of-year high-stakes state tests.

Best Practices in Academic Screening

Given the vast selection of reading screening assessments available, Albers and Kettler (2014) highlight important considerations for selecting an academic screening measure. First, it is necessary to determine the appropriateness of the screening measure such as what population is it being used for and how well it aligns with the particular domain of interest. Second, usability of a screener must also be considered including the resources, number of staff, and qualifications necessary to administer and score the screener. Amidst all of these things, most importantly, schools must also consider the technical adequacy of screening measures, including reliability, validity, and diagnostic accuracy of the scores the screening assessments provide.

The National Center on Intensive Intervention (NCII) defines reliability as the consistency of a set of scores that are designed to measure the same thing (2021). Internal consistency, or how well a set of item scores correlate with other item scores on the same test, is one type of reliability especially important to consider when selecting a screening measure. Validity refers to how well a screener measures what it is intended to measure and how well a score reflects the intended domain of interest or construct being assessed, which is a key component when being used to identify students at risk.

Diagnostic accuracy refers to how well the score on a screening assessment correctly identifies a student at risk versus not at risk for a negative future outcome. This is important for the accurate identification of students as either at-risk or not at-risk of failing the high-stakes state tests and is critical for school districts to consider as higher correlations should increase the classification accuracy of a screening measure.

While reliability and all forms of validity are foundational to a screening assessment's usefulness, validity with regard to classification accuracy is essential. Classification accuracy is related to how well a screening assessment predicts the future outcome that it is designed to assess. That is if a student's score classifies them as being at-risk at a screening assessment period, do they subsequently perform poorly on the high-stakes state test or the predicted outcome of interest? Likewise, if a student scores proficient on a screening assessment, do they perform within acceptable levels on a future construct of interest? Classification accuracy informs an assessment's diagnostic accuracy and the diagnostic accuracy of a screener is particularly important because it is used by school districts to make decisions about how to best use resources to support students in need of additional academic intervention (January & Klingbeil, 2020).

Screening measures with high diagnostic accuracy can accurately classify a student as at risk (true positive) or not at risk (true negative). Screening prediction errors then are incorrect classifications of either 1) being classified as at risk when a student is not truly at risk (false positives) or 2) failing to classify a student as at risk when they are truly at risk (false negative). Sensitivity, specificity, positive predictive value, and negative predictive value are ways to summarize and analyze screening assessment accuracy. Sensitivity is the proportion of truly at-risk students who were identified as at-

risk, that is the proportion of true positives on the outcome being predicted are correctly identified. Specificity is the proportion of all students who were truly not at risk of a problem who were correctly classified by the screener. The positive predictive value and negative predictive value are also important to consider when evaluating screeners. Positive predictive value is the proportion of students identified as at-risk who are truly at-risk while negative predictive value is the proportion of students identified as not at risk who are truly not at risk (Albers & Kettler, 2014).

Diagnostic accuracy indices are represented as scores that can range from 1.00 (perfect classification accuracy) to .00, where values at .50 indicate the predictor is no better than chance. Klingbeil, McComas, Burns, and Helman (2015) recommend that high sensitivity values (i.e., .90 or higher) are necessary for universal screeners, while moderate specificity values of .70 to .80 are acceptable. NCII rates a screening tool with its highest rating when it has a sensitivity rate of 70% or higher and a specificity rate of 80% or higher. In addition to individual values, sensitivity and specificity can be considered together when evaluating a screener through the area under the curve (AUC) statistical analysis. The AUC is an overall indication of the diagnostic accuracy of a screener. The NCII classifies screeners with an AUC of .80 or higher as having convincing evidence for overall diagnostic accuracy. This includes a sensitivity and specificity of .80 or higher. This information is necessary for schools to consider when determining which screening tools will best meet their needs given the resources available to them (NCII, 2021).

Curriculum Based Measurement as a Universal Screener

Curriculum based measurement (CBM) is a standardized assessment originally designed to assess students' growth in basic academic skills including reading, math, spelling, and written expression. The first CBMs developed assessed reading automaticity, or oral reading fluency (ORF), which measures how quickly and accurately a student can read. ORF assessments typically involve a student being timed while reading a grade-level passage and the examiner records the total number of words read minus the number of errors to get number of words read correctly. Strong performance on an ORF assessment requires the simultaneous integration of various reading skills to successfully read a short passage (Fuchs, 2004; Yeo, 2010). Although originally developed to measure academic progress in response to instruction, CBM and specifically ORF, has a wealth of research supporting its validity as an assessment of reading and a predictor of future reading achievement, such as performance on high stakes state tests (Deno, 2003; Hintz & Silbergliitt, 2005; Kettler & Albers, 2013; Shin & McMaster, 2019; Yeo, 2010).

Because of its ability to predict performance on high stakes state tests, ORF measures have become widely used as universal screeners in MTSS data-based decision-making practices, especially at the elementary level (Albers & Kettler, 2014; Yeo, 2010; Kettler & Albers, 2013; Shin & McMaster, 2019). Wayman, Wallace, Wiley, Ticha, & Espin (2007) conducted a synthesis of research literature on the usefulness of CBM in reading. Their review focused on studies related to the technical adequacy of CBM reading for instructional decision-making and progress monitoring. Given research at the time, they concluded that oral reading fluency (ORF) was consistently shown to be a better indicator of reading comprehension and reading proficiency than CBMs involving

maze selection and word identification, providing support for CBMs of ORF as markers for broader reading skills, not just reading speed.

Yeo (2010) completed a meta-analysis of 27 studies examining the relationship between CBMs of oral reading fluency and statewide achievement tests in reading for students grades 1 through 8. The studies included 19 different statewide achievement tests, five of which used the Minnesota Comprehensive Assessment (MCA). The studies were robust in terms of study sample size and included students grades 1 through 8. Across all studies, one-minute ORF passages used as a screener of reading skill were valid predictors of performance on the statewide reading tests examined. Studies with larger sample sizes had stronger correlation coefficients than did studies with smaller sample sizes. Further, correlations were somewhat lower when samples included larger proportions of ELL or special education students. Grade level and proportion of Caucasian, free/reduced-price lunch, or female students did not significantly impact the predictive validity between CBM and the statewide assessment. Results also indicated that the time between administration of the CBM and the statewide assessment was a significant predictor affecting the predictive validity with the more time between administrations yielding a lower correlation.

Kilgus, Methe, Maggin, and Tomasula (2014) also conducted a meta-analysis examining the use of curriculum-based measures of oral reading for universal screening purposes. Their study focused on the diagnostic accuracy of CBM and its ability to predict performance on criterion measures (i.e., statewide achievement tests and published norm-referenced tests). Their review included 34 studies with students in kindergarten to 8th grade who participated in universal screening using CBM on a tri-

annual basis (fall, winter, spring). Results replicated previous research demonstrating that CBM is a strong predictor of performance on future outcome measures. Results also indicated that CBM was able to successfully differentiate between those with and without reading problems. Some key findings were that the specific cut score used, along with the amount of time between the administration of the screener and the criterion measure, impacted the sensitivity and specificity, two key indicators of diagnostic accuracy. The results of this study support the use of CBM-R as a universal screener to determine the level of risk on a future outcome measure; however, the authors recommend local norms be used to determine a cut score that will provide diagnostic accuracy that is best suited for a school district given their needs and resources.

Computer Adaptive Tests as a Universal Screener

Computer adaptive tests (CAT) are increasingly being used for universal screening purposes. CATs are administered by computer and automatically adjust the difficulty of test items presented based on the test taker's performance. For example, a more difficult item is presented following a correct response and an easier item is presented following an incorrect response. CATs use the application of Item Response Theory (IRT) for real-time selection of test items such that each examinee is exposed to a minimum number of items beyond their ability level (Lu and Cong, 2016). From a feasibility standpoint, there are several advantages to CAT including better standardization of item presentation, easy administration with large groups of students, reduced costs associated with printing test materials, and improved security and data storage. In addition, results are available immediately allowing quick analyses of results (Clemens et al., 2015).

The NCII Tools Chart lists multiple CAT reading screeners with convincing evidence for screening uses given their strong correlations with performance on end-of-year state tests. Data reviewed on the website is based on information provided by the various assessments' authors. There is less published research on the usefulness of CATs as academic screeners, however.

Ball and O'Connor (2016) examined the predictive validity of the Measures of Academic Progress (MAP) for performance on the Wisconsin Knowledge and Concepts Exam (WKCE). The Reading MAP and Language Usage MAP are both CATs aligned to the state accountability standards. The Reading MAP (R-MAP) is a CAT that evaluates the reading skills related to word analysis and vocabulary, basic understanding of text, and analyzing and extending meaning. The Language Usage MAP (L-MAP), is a CAT that evaluates language skills related to developing writing, parts of speech, and capitalization and punctuation. This study specifically examined how well R-MAP and L-MAP predicted future reading performance when used alone or in combination with an oral reading fluency screener in the spring of second grade. Overall results demonstrated that MAP and ORF scores from the end of second grade correctly classified students as at-risk or not at-risk of poor performance on the third-grade WKCE. In regards to both predictive validity and classification accuracy, L-MAP emerged as the most robust single predictor of WKCE performance.

National Center for Intensive Intervention (NCII) Review

The National Center for Intensive Intervention (NCII) is a website that provides ratings of screening assessments to allow consumers to make informed decisions regarding what specific tools will best meet their needs. Assessment authors and

publishing companies submit materials to be reviewed during a call for submissions period. The tools charts provide expert ratings on the technical rigor of screeners submitted for review. Ratings are determined by an external Technical Review Committee using established criteria. Criteria include classification accuracy, technical standards, and usability features. Screening assessments' reliability, validity, and classification accuracy (fall, winter, and spring) are classified as being in one of four categories: convincing evidence, partially convincing evidence, unconvincing evidence, or data unavailable. Usability features examined include the administration format, administration and scoring time, scoring format, types of decision rules, and evidence available for multiple decision rules.

FastBridge's CBM-Reading is a Curriculum-Based Measurement of oral reading fluency reviewed on the NCII Academic Screening Tools Chart (2021). Usability-wise, this is an individually administered screener that takes about 3 minutes and has automatic score uploading when web-based scoring is used. The validity of CBM-R is rated with convincing evidence from grades 1 through 6. Evidence for CBM-R's concurrent and predictive validity on the NCII website include correlations with AIMSweb ORF scores for a sample of students from Minnesota with approximately 220 students per grade for grades 1-6. Partially convincing validity ratings were given for 7th and 8th grade because some correlations did not meet the expected $r = .60$. The criterion measure for grade 7 and 8's concurrent and predictive validity were NWEA's MAP reading assessment, which is a broad indicator of overall reading ability rather than an ORF measure. The reliability for CBM-R is given the rating of convincing evidence for all of 1st through 8th

grade, which suggests at least two forms of reliability, alternate-form and inter-rater, met or exceeded the expected 0.70.

When examining the classification accuracy for CBM-R, 2nd grade is rated with convincing evidence for fall, winter, and spring. Convincing evidence indicates an area under the curve (AUC) score, sensitivity and specificity are all greater than or equal to 0.80. This suggests CBM-R is an accurate tool for use as a screening assessment in 2nd grade. Classification accuracy evidence is less consistent across the year for grades 1, 3, 4, 5, 6, 7, and 8 with evidence ranging from convincing to partially convincing. Partially convincing evidence indicates an AUC greater than or equal to 0.70 but lower than 0.80 or sensitivity and specificity greater than or equal to 0.70. Classification accuracy for the fall of 1st grade was rated as having unconvincing evidence. This means that the AUC, sensitivity and specificity were below 0.70.

FastBridge's aReading is also reviewed on NCII's tools chart (2021). aReading is a computer-adaptive test (CAT) of broad reading ability used for universal screening throughout the academic year for students in Kindergarten through 12th grade. Usability-wise, the screener can be administered at the individual, small, or large group level and takes an average of 23 minutes to complete. Validity for aReading is rated with convincing evidence for grades 2 through 8, but only partially convincing evidence for grade 1 indicating the research was mixed and did not meet the expected 0.60. Evidence for aReading's predictive validity on the NCII website include correlations with the Gates MacGinitie Reading Tests-4th Edition (GMRT-4th) for 1st and 2nd grade from 2 schools in Minnesota with 55 to 125 students in 1st grade and 215 to 300 students in 2nd grade. Concurrent and predictive validity for grades 3 through 8 were evidenced by scores on

Georgia Milestones Assessment System (Georgia Milestones) from all students completing the FastBridge aReading for universal screening in the fall (predictive) and spring (concurrent) and students completing the Georgia Milestones assessment in the spring with approximately 8,000 to 12,000 student scores per grade and season.

Reliability for aReading is rated as having convincing evidence for all of kindergarten through 8th grade, suggesting at least two forms of reliability, test-retest and IRT-score based, met or exceeded the expected 0.70. The classification accuracy for aReading is rated with convincing evidence for fall, winter, and spring for all of 2nd grade through 8th grade and the spring of 1st grade. Data was unavailable for the classification accuracy of aReading for the fall, winter, and spring of kindergarten and fall and winter of 1st grade. Classification accuracy data is based on aReading's ability to predict future performance on a criterion measure, Georgia Milestones Assessment System, which begins being administered in the spring of 3rd grade.

Statement of the Problem

Accurately predicting students at risk for reading difficulties and poor performance on future high-stakes state testing is necessary for making decisions regarding who should receive intervention. Being able to identify the students who are at risk with as few resources (i.e., staff, time, expense) as possible is also important for schools. Further, while large research studies and the NCII website provide important guidance for schools to evaluate and select screening assessments, local analysis of screening accuracy is recommended. The purpose of the current study is to examine the diagnostic accuracy and predictive validity of FastBridge's CBM-R and aReading assessments as reading screening assessments for identifying 2nd grade students who are

at-risk to not show adequate reading proficiency on the Reading Minnesota Comprehensive Assessment – Third Edition (Reading MCA-III; Minnesota Department of Education, 2022). This research will allow more information for educators regarding which screener is the better predictor of students correctly classified as at risk or not at risk at the elementary level.

The research question addressed is:

- 1) Which fall screener, CBM-R or aReading, is more accurate at predicting 2nd grade students who will be classified as “not proficient” on the Reading Minnesota Comprehensive Assessment Series-III (MCA-III) taken in 3rd grade?

Chapter 2

Methods

This chapter describes the methods of the study. First, participant demographics and a description of the education 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

FastBridge academic screening data was obtained for 2nd grade students from the 2017-2018 academic year who also had scores obtained for the Minnesota Comprehensive Assessment – Third Edition (MCA-III) for 3rd grade of the 2018-2019 academic year from a rural school district in the upper Midwest. There was a total of 185 students who had taken both the CBM-R and aReading assessments in the fall of 2017. Of the 185 students who had screening scores from 2nd grade. 164 students also completed the MCA-III in their 3rd grade year. There were no known factors contributing to the difference in numbers between 2nd and 3rd grade, rather the discrepancy is likely due to common factors such as students moving in and out of district or families opting their student out of taking the MCA-III assessment. A majority of the 164 students in the final sample identified as White at 97.6%, 1.2% Black or African American, and >1% identified as Hispanic/Latino. The sample was approximately half female at 48.8%. Additionally, 6.7% received special education services, 1.8% of the sample were English Language Learners, and 36.6% were eligible for free or reduced lunch. No other demographic data was provided by the school district.

Table 1

Sample Characteristics (N=164)

Characteristic	Percent of Student Sample
White/Nonwhite	97.6% / 2.4%
Female/Male	48.8% / 51.2%
Special Education	6.7%
ELL	1.8%
FRL	36.6%

Note: FRL = Free or Reduced Lunch; ELL = English Language Learners

Measures***FastBridge CBMReading (CBM-R)***

FastBridge CBMReading (CBM-R; Illuminate Education, 2022) is a universal screening assessment for reading in grades 1 through 8. Text type, paragraph and sentence structure, word and language usage, and cohesion were key features addressed in the development of all FastBridge CBM-Reading passages. The passages were designed with detailed specifications and in consultation with educators and content experts. Passages were originally tested with at least 500 students per grade level. Researchers analyzed data from trials of the passages and edited them to optimize the semantic, syntactic, and cultural elements.

CBM-R is a one-minute assessment of oral reading fluency. Students read three short passages written at a 2nd grade level when used for screening in 2nd grade. Examiners score word reading errors according to standardized procedures. Errors include omission, insertions, substitutions, and mispronunciations (Illuminate Education, 2022). A student's performance on each passage results in four scores: total words read, number of errors, words read correctly per minute (total words read minus errors), and

percent of words read correctly. The median words read correctly from the three screening passages is used as the overall screening score.

The CBM-R technical manual (FAST Technical Manual, 2018) indicates that students who score at or above the 40th percentile of a national sample of 2nd grade students are likely to later score in the proficient range on subsequent state standards assessments. This means that 2nd grade students whose oral reading fluency scores were above 58 words read correct per minute were classified as showing minimal risk for future reading challenges. Scores below 58 WRCM were in the some-risk range and scores below 30 WRCM were in the high-risk range. High-risk status corresponds with scores at or below the 20th percentile on a national sample. For the current study, screener results were dichotomized dividing it into “at-risk” which includes the high-risk and some-risk levels (<40th percentile) and “not at-risk” which includes the low-risk level (\geq 40th percentile). These percentiles of scores are relative to same-grade peer performance nationally.

NCII (2021) lists 2nd grade CBM-R alternate-form reliability as .90 and its inter-rater reliability as .97. The concurrent validity for CBM-R with AIMSweb Oral Reading Fluency scores in 2nd grade is .97 while the predictive validity for CBM-R, which was determined using AIMSweb Oral Reading Fluency as the criterion measure is .92 indicating strong evidence for its use as a screener used to identify a student’s risk status.

FastBridge Adaptive Reading (aReading)

FastBridge Adaptive Reading (Illuminate Education, 2022) is a computer-adaptive measure of broad reading ability. It can be used for screening with students in Kindergarten through 12th grade. This untimed assessment is designed to align with the

national Common Core Standards. Items on aReading for students in Kindergarten through 5th grade were designed to assess concepts of print, phonological awareness, phonics, vocabulary, and comprehension. aReading can be administered individually or in a small or large group setting using a computer. Instructions are provided via the online program. Administration time averages 23 minutes per student or group (NCII, 2021).

The FastBridge Technical Manual (2018) reports screening accuracy data for aReading when used to predict proficiency on the Gates MacGinitie Reading Tests-4th Edition (GMRT-4th) and Measures of Academic Progress (MAP) with the MCA-III assessment. NCII (2021) lists 2nd grade aReading IRT-score based internal consistency reliability as .96 and its test-retest reliability as .90. The construct validity correlation for aReading with MAP in 2nd grade is .83 while the predictive validity with the GMRT-4th is .75 indicates strong evidence for its use as a screener to identify students' risk status.

The scores generated by aReading are based on an Item Response Theory (IRT) logit scale similar to other educational measures. Scores on aReading have a lower bound of 350 and an upper bound of 650 with a middle value of 500 and a standard deviation of 50. Like CBM-R, students whose screening scores fall at or above the 40th percentile when compared to a national sample of children in the same grade are considered to be at low-risk for future reading difficulties.

Minnesota Comprehensive Assessment – Third Edition (MCA-III)

The Minnesota Comprehensive Assessment – Third Edition (Minnesota Department of Education, 2022) is a comprehensive state test that assesses student proficiency with academic skills and knowledge identified in the Minnesota academic

standards. The MCA-III assesses reading, mathematics, and science. The Reading MCA-III is administered to students in 3rd through 8th and 10th grade. The group administered assessment includes 48 multiple choice items organized under two substrands: literature and informational text. Items may assess students' ability to analyze, interpret and evaluate fiction and nonfiction works (MCA Technical Manual, 2020).

The MCA-III is designed such that all 3rd grade students will complete similar test forms. Students in 3rd grade are given four to seven passages and 48 items (Technical Manual for Minnesota's MCA and MTAS assessments, 2020). For the MCA-III, scores range from 1 to 99 and are prefixed by the student's grade (i.e., g1 to g99 where g is the students' grade). For example, 3rd grade student scores range from 301 to 399.

MCA-III scores are divided into four categories of proficiency: Does Not Meet the Standards (scores below 340), Partially Meets the Standards (340-349), Meets the Standards (350 to the next cut score), and Exceeds the Standards (374-399). For the current study, MCA-III scaled scores were divided into a dichotomous "proficient" which included Meets the Standards and Exceeds the Standards and "not proficient" which includes Does Not Meet the Standards and Partially Meets the Standards categories. This specifically separated scaled scores such that scores of 301-349 were considered not proficient and scales scores of 350-399 were considered proficient.

Procedure

Per school district policy, all fall academic screeners were administered between September and October. The classroom teacher proctored the aReading assessment in a large group setting on individual computers. Students who were absent on the day of

screening completed the test individually or in a small group upon their return to school. Students were also administered three CBM-R probes with a trained staff member in a 1:1 setting. Probes were hand scored and the median score of words read correct in one minute was reported for each student.

The MCA-III is typically completed in April 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 due to severe special education impairment.

IRB approval from the University of Wisconsin-Eau Claire was obtained prior to the release of data from the school district. Data was obtained archivally from the school district. All identifying information for students was removed prior to being shared with the researcher. To maintain data security, the data files were password protected.

Data Analysis

Descriptive Statistics

Descriptive statistics analyzing the mean, standard deviation, skewness, and kurtosis were completed to analyze score distributions of the measures. Pearson correlations were completed between aReading and CBM-R, aReading and MCA-III, and CBM-R and MCA-III to compare the similarities of scores.

Dichotomous Classification

All measures used for the current study classify student scores into multiple levels of risk or proficiency. Thus, for the purposes of this study, all assessments were dichotomized at the onset of data analysis. Both FastBridge tools, CBM-R and aReading, are divided into four levels of risk: High-Risk, Some-Risk, Low-Risk, and Exceeds. For the current study, both screeners were dichotomized dividing them into “at-risk” which includes the High-Risk and Some-Risk levels and “not at-risk” which includes the Low-Risk and Exceeds levels. For the current study, MCA-III scores were divided into a dichotomous “proficient” which includes Meets the Standards and Exceeds the Standards, and “not proficient” which includes Does Not Meet the Standards and Partially Meets the Standards categories. Specifically, spring MCA-III scores at or below the 43rd percentile were considered not proficient and scores at or above the 57th percentile were considered proficient.

Diagnostic Accuracy

Cross tabs analyses and hand calculations of classification accuracy statistics were used to assess the sensitivity, specificity, positive predictive power, and negative predictive power of CBM-R and aReading.

Area Under the Curve

Receiver operating characteristics (ROC) curves were used to inspect the diagnostic accuracy of each screening assessment. Analyses were conducted using SPSS V27.0 (IBM CORP., 2020). The ROC analysis includes measures of sensitivity and specificity to determine the area under the curve (AUC; Compton, et al, 2006; Klingbeil et al., 2015). In this study, the ROC curves plotted true positives against false positives on the MCA-III. AUC statistics can range from poor diagnostic abilities, 0.5, to perfect

diagnostic abilities, 1.0 (Compton, et al., 2006; Klingbeil et al., 2015). As suggested by Compton et al. (2006), when assessing a tool's diagnostic ability, ROC curve values $>.90$ are considered excellent, while values from $.80$ to $.90$ are good, $.70$ to $.80$ are fair, and below $.70$ are poor.

Chapter 3

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. Students with missing data were excluded therefore there were no alterations required for CBM-R or aReading scores. The mean performance in the fall for 2nd grade students on CBM-R was 62 words read correct per minute (range = 2 - 161). The mean performance for 2nd grade students in the fall for aReading was 475 (range 381 - 516). On the fall screening assessment, approximately 60% of students performed in the proficient range on CBM-R and 70% performed in the proficient range on aReading. In the spring, third grade, approximately 59% of students scored in the proficient range on the MCA-III Reading assessment. The mean MCA-III performance was a scaled score of 353 with scores ranging from 311-392. Finally, high correlations were observed between CBM-R and aReading ($r=.84$) and correlations for CBM-R and MCA ($r=.65$) and for aReading and MCA ($r=.76$) were moderate. These values are lower than the ranges of screener to state test correlations (CBM-R and MCA $r = .76$; aReading and MCA $r = .82$) as reported by the test publishers (FAST Technical Manual, 2018)

Table 2

Descriptive Statistics for Reading Screeners and State Test (N = 164)

Assessment	<i>M</i>	<i>SD</i>	Skewness	Kurtosis	Not Proficient
Fall CBM-R	62.21	33.52	0.37	-0.05	40.2%
Fall aReading	474.82	21.49	-0.83	1.78	29.9%
Spring Reading MCA-III	353.38	17.51	-0.08	-0.27	41.5%

Note: CBM-R = FastBridge Curriculum Based Measurement-Reading; aReading = FastBridge Adaptive Reading; Reading MCA-III = Minnesota Comprehensive Assessment-Series III in Reading

Table 3

Correlation Between Variables

Assessment	<i>r</i>
CBM-R and aReading	.84
CBM-R and MCA-III	.65
aReading and MCA-III	.76

Research Question

This study examined the diagnostic accuracy of the fall administration of each individual 2nd grade screener, FastBridge CBM-R and FastBridge aReading, in predicting performance on the Minnesota Comprehensive Assessment Series-III taken in the spring of 3rd grade. CBM-R screening resulted in fair specificity of .78 and a sensitivity value of .66 which is considered poor. aReading screening resulted in excellent specificity of .93; however, the sensitivity was poor at .62. Poor sensitivity values indicate many students were misclassified as “not at risk” during screening but went on to perform below proficiency on the MCA-III. Further, when examining the positive predictive values (PPV) for each measure, CBM-R demonstrated a low value of .68, while aReading demonstrated good PPV at .86. Both CBM-R and aReading demonstrated fair negative predictive values (high rates of false negatives) with a value of .77 for both measures.

Both screeners’ scores demonstrated good diagnostic accuracy (AUC .89, .83 respectively). Compton et al. (2006) suggests classifying AUC values as follows: >.90

are excellent, values ranging from .80 -.90 are good, .70 to .80 are considered fair, and values below .70 are poor.

Table 4

Diagnostic Accuracy of aReading and R-CBM Universal Screening measures (N=164)

Assessments	TP	TN	FP	FN	AUC	Sensitivity	Specificity	PPV	NPV
Fall CBM-R	45	75	21	23	.83	.66	.78	.68	.77
Fall aReading	42	89	7	26	.89	.62	.93	.86	.77

Note. aReading = Adaptive Reading; CBM-R = Reading Curriculum Based Measurement; aReading+CBM-R = Combination of Adaptive Reading 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 = $TN/(TN+FN)$.

Chapter 4

Discussion

The final chapter discusses the study's findings and implications. Specifically, it discusses the implications for reading screeners used at the elementary level and the predictive validity of CBM-R and aReading screening assessments. Next, implications for practice are discussed. Finally, recommendations for potential future research and limitations of the study are reviewed.

Universal academic screening is a key component in the identification of students in need of early intervention within a RtI or MTSS framework (Jenkins, Hudson, & Johnson, 2007; Klingbeil, McComas, Burns, & Helman, 2015; Speece et al., 2011). This study compared the diagnostic accuracy of two different screening measures, one curriculum-based measure (CBM) of oral reading fluency (ORF) and one computer-adaptative test (CAT) of reading for a local sample of 2nd grade students in one specific school district. The focus was to determine which screener, when taken in the fall of 2nd grade, more accurately predicted students who were classified at-risk to perform in the “not proficient” range on the end-of-year high stakes assessment in the following academic year.

Overall, both screeners had good diagnostic accuracy when the AUC values were considered. AUC statistics are determined from Receiver Operator Curve analysis, which are used to determine a specific cut-scores that will maximize sensitivity and specificity to yield the strongest diagnostic accuracy (Klingbeil et al., 2015; Silbergliitt & Hintze, 2005). AUC statistics range from poor diagnostic accuracy, .50 which is no better than chance, to 1.0 which is a perfect predictor between students who are at-risk

and not at-risk (Compton, et al., 2006; Klingbeil et al., 2015; Speece et al., 2011). aReading's AUC value of .89 was slightly higher than CBM-R with an AUC of .83. The diagnostic accuracy scores are both considered to be "convincing evidence" according to the NCII screening tools chart. These results indicate that for this sample of students, fall of 2nd grade aReading scores correctly predicted students' risk level on the MCA-III assessment administered in the spring of 3rd grade with 89% accuracy while fall of 2nd grade CBM-R was able to correctly classify risk status with 83% accuracy.

Despite the overall accuracy of both screeners being in the good range, when the vendor-provided cut scores were used to identify prediction errors, levels of sensitivity for both screeners were more concerning. Both CBM-R and aReading demonstrated moderate levels of sensitivity meaning the measures incorrectly classified students as at risk or failed to identify many students who subsequently went on to perform below expectations on the MCA-III. CBM-R and aReading had similar false negative rates (23 and 26 respectively). This is problematic because when students are not accurately identified as being at risk during screening, they may not be considered for intervention that could help prevent more intractable reading challenges in the future (Compton et al, 2006; Klingbeil et al., 2015). When considering this within a MTSS framework, at the Tier 1 level, universal screening is a key component that assists in the early identification of those who need additional supports or intervention (Albers & Kettler, 2014; Stoiber, 2014). Universal screeners are also used to evaluate the overall effectiveness of programs at the Tier 1 level (Albers & Kettler, 2014). When screeners are not accurate in their ability to identify those at-risk of future academic difficulties, the ability to use this data to determine the true effectiveness of Tier 1 reading programs is inhibited. This

further highlights the importance of universal screening measures' ability to be accurate in the identification of students as at-risk of academic problems.

When examining specificity rates, aReading outperformed CBM-R. aReading demonstrated excellent specificity while CBM-R's specificity, while in the acceptable range, was lower in its ability to accurately identify students as not at-risk in the fall of 2nd grade. These results indicate that, when taken in the fall of 2nd grade, aReading was overall better than CBM-R at identifying students as not at-risk for those who ultimately passed the MCA-III in the spring of 3rd grade. When examining the number of false positives, aReading had 7 while CBM-R had 21 students who were incorrectly identified as at-risk on the screeners and went on to pass the MCA-III in the spring of 3rd grade. Although in the acceptable range, CBM-R's specificity of .78 also means 22% of students were classified as needing intervention who actually went on to meet 3rd grade standards for reading on the MCA. Lower specificity values can be problematic for multiple reasons including school districts may use valuable resources for students who ultimately don't need them or students receiving additional supports when they don't need them which may limit them from other opportunities.

When examining the false positives, at first it may seem that the numbers suggest students were incorrectly classified as being at-risk and went on to pass the MCA-III. Another consideration is that those students who were classified as being at-risk on the fall screener, may have been provided some intervention. These false positive may actually be the result of students who were classified as at-risk, receiving a successful which ultimately resulted in those students passing the MCA-III.

The positive predictive values for each screener provide one illustration of why the specificity values varied across the two measures. Positive predictive value (PPV) refers to the proportion of students who were identified as needing help and actually did need help (Albers & Kettler, 2014). Using the author-defined cut scores for some risk and high risk, the PPV of aReading was .86 suggesting 86% of students identified as at-risk in the fall of 2nd grade actually went on to perform in the not proficient range on the MCA-III assessment in 3rd grade. However, CBM-R had a PPV of .68 meaning that more than 30% of students identified as at-risk in the fall of 2nd grade went on to score in the proficient range on the 3rd grade MCA-III. The negative predictive value (NPV) was in the moderate range for both aReading and CBM-R. This suggests both screeners are equally accurate at identifying those who do not need additional supports and both screeners missed a large number of students who went on to fail the MCA-III assessment in the spring of 3rd grade.

When post hoc analyses were calculated removing the “some-risk” group of students from the data, similar patterns resulted. Specificity values increased for both screeners with aReading’s specificity being .98 and CBM-R’s .94. However, the removal of the ‘some risk’ classifications further decreased sensitivity values to .55 for CBM-R and only .42 for aReading. These results indicate that when this group of students is removed, the number of false negatives increases while the number of false positive decreases for both screeners. In other words, the number of students who were not identified as “at-risk” on the screener and go on to perform in the “not proficient” range on the MCA-III increases when the “some-risk” group of students is removed from the data. There were 26 students who performed in the “not at-risk” category on aReading

and subsequently went on to perform in the “not proficient” range (false negative) on the MCA-III while CBM-R had 23 false negatives.

Implications for Practice

The results discussed above suggest some implications for practice. First, this study demonstrated that both screeners, CBM-R and aReading, had good AUC scores, meaning they both appear to be strong screeners with this particular sample of students for use within a MTSS framework. Given the diagnostic accuracy is similar for both screeners, it is important for school districts to consider the usefulness of administering both screeners three times a year. It may be beneficial for schools to consider the pros and cons of each screener to determine which is most effective for their setting. For example, CBM-R although brief to administer individually, when administered to large groups will ultimately require a lot of instructional time and it may therefore be more reasonable to use aReading which can be administered to large numbers of students at once. aReading requires technology that CBM-R does not and this too is a relevant factor for schools to consider. The errors patterns are another important factor to consider when determining which screeners to administer. Given that the errors patterns differed between the two screeners, it may be beneficial for the school district to administer both if it ultimately leads to more accurate classifications. Another option would be to consider a gated, or multiple-gate, approach to screening. Gated screening consists of conducting more tests on fewer students to increase the accuracy of identifying students at-risk (Albers & Kettler, 2014). For example, since aReading demonstrated overall better diagnostic accuracy, it may be used for the universal screening measure for the initial identification of students who are at-risk. The schools may then choose to use CBM-R

with the students who were identified by aReading or they may choose use some other data to help inform their decision-making process in identifying the truly at-risk students.

A final implication for practice relates to the use of vendor-determined cut scores for identifying students at risk for future reading challenges. Even though overall AUC scores were positive, there were challenges to sensitivity for aReading and CBM-R when nationally-determined cut scores were used to classify this local sample of students as ‘high risk’ and ‘some risk.’ Both screeners missed a number of students who ultimately performed in the “not proficient” range on the MCA-III. Thus, this study provides one more illustration of the importance of evaluating vendor-recommended cut scores when used locally. Schools may benefit from considering trends in screener data for predicting performance on state tests and determining cut scores based on local data. Some research suggests that the preceding year’s test performance is the best way to determine cut-scores (Klingbeil, Osman, Carrigan, Paly, & Berry-Corie, 2021; VanDerHeyden, 2013). Further, some research suggests that using local cut-scores derived from prior year statewide achievement tests can improve sensitivity and still maintain adequate specificity (Klingbeil et al., 2021). For schools lacking resources to complete these types of analyses with local data, the findings illustrate the importance of considering multiple pieces of data when determining student intervention needs in an MTSS system. However, differences with the cut-score values were found therefore practitioners should consider data locally and look for trends in how the screeners’ accuracy for different cut-scores or predictions may vary. Multiple pieces of data should always be used to inform decision-making.

Implications for Future Research

The discussion of results of this study lead to some implications for future research to consider. This study examined the diagnostic accuracy of CBM-R and aReading using the fall benchmark in 2nd grade to predict the performance on the MCA-III in the spring of 3rd grade. Research suggests that when the data used from screeners with less time between their administration and the administration of the criterion measure, diagnostic accuracy is improved (Ball & O'Connor, 2016; Kilgus et al., 2014). Future research may want to examine different screening benchmarks closer to the criterion measure.

Second, future research should examine the diagnostic accuracy of multiple screening measures used simultaneously. The current study examined the CBM-R and aReading universal screening tools' diagnostic accuracy individually. The school district used in this study administer both of these tools during each screening period, three times per academic year. It would be valuable for future research to examine the diagnostic accuracy of the screener combined rather than individually.

Last, future research could explore alternative cut scores or easily replicated strategies for schools to follow to refine the accuracy of decision-rules when screeners are used to identify students in need of additional intervention.

Limitations

Results of the current study should be interpreted in light of possible limitations. One limitation relates to amount of time between the administration of the screening measures and the outcome measure. The selection of this data was intentional to examine the diagnostic accuracy of the screeners at a time when early intervention can be provided to prevent more persistent reading challenges. Research supports that diagnostic

accuracy is stronger when the administration of the screener is closer to the administration of the outcome measure (Kilgus, et al., 2014). Thus, results for this local sample of students might be more like published data if the timeframe of screening and outcome measure analysis replicated vendor reported data. Meaningful assessments administered at the end of 2nd grade could also be used as an outcome measure in order to help determine screener accuracy.

A second limitation for this study is related to how the screener risk status and MCA proficiency levels were dichotomized. CBM-R and aReading screeners provide four classifications of risk status: high risk, some risk, low risk, and exceeds. For this study, the “some-risk” screener classifications were combined with the “at-risk” classifications, and “low risk” was combined with “exceeds”. This dichotomization allows school districts to analyze data more easily. However, it introduces challenges when interpreting sensitivity values. When the data is divided this way, students who perform slightly below the risk threshold are categorized with those who perform well below the risk threshold therefore placing all of these students at the same level of risk despite differences in each student’s true level of risk (Klingbeil, Van Norman, Nelson, & Birr, 2019). For the current study, the CBM-R data illustrated this limitation. The students identified as “some-risk” split almost evenly with 17 going on the perform in the “not proficient” range and 16 going on the perform in the proficient range on the MCA-III.

Last, student attrition from 2nd to 3rd grade may have had unknown effects on the current sample. While there were no known factors contributing to the differences in student enrollment and test score availability between 2nd and 3rd grade, 21 students who

completed screening assessments in 2nd grade were not represented with MCA-III scores in 3rd grade. While it was assumed that the discrepancies were due to transiency or families opting out of taking the MCA-III assessment, there may have other factors and the factors may have an unknown association with screening score or MCA III performance that limits interpretations of the resulting data. Further, this study did not analyze whether attrition was equally distributed across screener risk levels.

Conclusion

Universal screening is a widely used practice to determine which students are at risk for poor performance on high stakes state tests and therefore not meeting grade level standards. The results of the current study demonstrate that aReading and CBM-R had similar results but aReading was the overall better screener when assessing diagnostic accuracy. Results did indicate a number of errors in the correct classification of students as either at-risk or not at-risk using vendor-provided cut scores and demonstrate the importance of considering local norms for the most accurate screening decisions. Given the diagnostic accuracy was similar for both assessments, results suggest that schools may want to consider selecting one screening measure depending on which tool is most effective for their unique setting. School should also consider the use of additional data or multiple-gate screening to improve classification accuracy. Additional research is needed to examine the diagnostic accuracy of both screeners combined, using additional data to inform decision making, or exploring different cut-scores for defining “at-risk” students.

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 121-132). Bethesda, MD: National Association of School Psychologists.
- Ball, C. R., & O'Connor, E. (2016). Predictive utility and classification accuracy of oral reading fluency and the measure of academic progress for the Wisconsin Knowledge and Concepts Exam. *Assessments for Effective Intervention, 41*(4), 195-208. doi: 10.1177/1534508415620107
- Castillo, J. M. (2014). Best Practices in Program Evaluation in a Model of Response to Intervention/Multitiered System of Supports. In Harrison, P.L. & Thomas, A. (Eds.), *Best Practices in school psychology: Foundations* (pp 329-342). Bethesda, MD: National Association of School Psychologists.
- Center for Response to Intervention. Retrieved November 28, 2021 from <https://mtss4success.org/>
- 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.
- 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.

- Deno, S. L. (1985). Curriculum-Based Measurement: The Emerging Alternative. *Exceptional Children, 52*, 19-232.
- Deno, S. L. (2003). Developments in Curriculum-Based Measurement. *The Journal of Special Education, 37*(3), 184-192.
- Every Student Succeeds Act, 20 U.S.C. § 6301. (2015).
- Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*(2), 188-192.
- Hintz J. M. & Silbergliitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review, 34*(3), 372-386.
- 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.
- Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*(4), 582-600.
- 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.
- 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*(2014), 377-405.

- 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 School, 52*(5), 500-514.
- Klingbeil, D. A., Osman, D. J., Carrigan, J. E., Paly, B. J., & Berry-Corie, K. (2021). Evaluating aimswebPlus Math as a universal screening measure in upper elementary and middle school. *American Psychological Association, 36*(2), 97-106.
- Klingbeil, D. A., Van Norman, E. R., Nelson, P. M., & Birr, C. (2019). Interval likelihood ratios: Applications for gated screening in schools. *Journal of School Psychology, 76*, 107-123.
- Lu, P. & Cong, X. (2016). The research on computerized adaptive testing. *Journal of Physics: Conference Series, 710*, 1-10. doi:10.1088/1742-6596/710/1/012029
- Minnesota Department of Education. (2022). *Statewide Testing*.
<https://education.mn.gov/MDE/fam/tests/index.htm>.
- National Center on Intensive Intervention. (2021, July). *Academic screening tools chart*.
<https://charts.intensiveintervention.org/ascreening>.
- Pearson Inc. (2020). 2019-2020 Technical manual for Minnesota's MCA and MTAS assessments. Minnesota Department of Education. Minneapolis, MN.
- Proctor, S. L., & Meyers, J. (2014). Best Practices in Primary Prevention in Diverse Schools and Communities. In Harrison, P. L. & Thomas, A. (Eds.), *Best Practices in School Psychology: Foundations* (pp 33-47). Bethesda, MD: National Association of School Psychologists.

- 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. U.S. Department of Education 2019).
- Shapiro, E. S., & Gebhardt, S. N. (2012). Comparing computer-adaptive and curriculum-based measurement methods of assessment. *School Psychology Review, 4*(3), 295-305.
- Shin, J. & McMaster, K. (2019). Relations between CBM (oral reading and maze) and reading comprehension on state achievement tests: A meta-analysis. *Journal of School Psychology, 73*, 131-149.
- 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. doi: 10.1177/073428290502300402
- Speece, D. L., Schatschneider, C., Silverman, R., Case, L. P., Cooper, D. H., & Jacobs, D. M. (2011). Identification of reading problems in first grade within a response-to-intervention framework. *The Elementary School Journal, 111*(4), 585-607.
- Stoiber, K. C. (2014). A comprehensive framework for multitiered systems of support in school psychology. In Harrison, P. L. & Thomas, A. (Eds.), *Best Practices in school Psychology: Data-Based and Collaborative Decision Making* (pp 41-70). Bethesda, MD: National Association of School Psychologists.

- Theodore J. Christ and Colleagues (2018). Formative Assessment System for Teachers™ Technical Manual, Minneapolis, MN: Author and FastBridge Learning.
- VanDerHeyden, A. M. (2013). Universal screening may not be for everyone: Using a threshold model as a smarter way to determine risk. *School Psychology Review*, 42(4), 402-414.
- Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education*, 41(2), 85-120.
- 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. doi: 10.1177/0741932508327463