August 2021

Using Cost-effectiveness Analysis to Select Mathematics Screening Measures in Middle School

Samuel Maurice
University of Wisconsin-Milwaukee

USING COST-EFFECTIVENESS ANALYSIS TO SELECT MATHEMATICS SCREENING MEASURES IN MIDDLE SCHOOL

by

Samuel A. Maurice

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Educational Psychology

at

The University of Wisconsin – Milwaukee

August 2021

ABSTRACT

USING COST-EFFECTIVENESS ANALYSIS TO SELECT MATHEMATICS SCREENING MEASURES IN MIDDLE SCHOOL

by

Samuel A. Maurice

The University of Wisconsin – Milwaukee, 2021

Under the Supervision of Professor David Klingbeil, Ph.D.

This study examined the utility of using cost-effectiveness analysis to select universal mathematics screening measures in middle school. Participants (*n*=1586) were students in Grades 6, 7, and 8 at two suburban middle schools in Wisconsin. Screening data, including previous year criterion-measure (Wisconsin Forward Exam) scores, fall Measures of Academic Progress scores, and curriculum-based measurement scores were collected in the fall of 2016. Multiple imputation was used to account for missingness, and linear combinations of screening scores were created using receiver operator curve analyses. Costs were calculated based on published standards, the CostOut® Toolkit, and the experience of content experts. Results reveal that the single most cost-effective screening method studied was using students' previous year criterion-measure scores to predict current year risk. The most cost-effective linear combination of screening methods was the Wisconsin Forward Exam and Measures of Academic Progress. An analysis of coefficients of variance revealed that using cost-effectiveness analysis produced more variability among screening methods than when using diagnostic accuracy alone, potentially helping stakeholders select from among multiple screening approaches. Finally, the results of this study were tested for robustness to changes in cost assumptions. Analyses revealed

that the results of this study were very robust, even when costs were changed significantly. Implications of this study suggest that cost-effectiveness analysis could prove useful in selecting universal academic screening measures, that schools and districts may be able to utilize criterion-measure data in place of other screening approaches, and that combinations of screening measures, although more expensive than individual measures, may indeed be more cost-effective.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

John Donne wrote that no man is an island. It is my deepest conviction that no person achieves success solely through their own effort. To all my mentors, friends, and loved ones who taught me, inspired me, and cared for me, this accomplishment is yours as much as mine.

For my former advisor and dissertation chair, Dr. David Klingbeil. Dr. Klingbeil gave me his time, his effort, and his advice every time I needed them for the past five years. His support throughout graduate school and in completing this dissertation has been invaluable. Without his guidance, this dissertation would not have happened.

For Dr. Kyongboon Kwon, my advisor for the past three years. Dr. Kwon's compassion and professionalism have been an inspiration for me throughout graduate school. I hope to someday become the type of psychologist, advisor, and teacher that she is.

For Dr. Razia Azen, my committee member and former boss. Dr. Azen's superior intellect, constant encouragement, and boundless patience reinforced my passion for statistics and education. She was always available to provide mentorship, friendship, and to discuss our shared passion, hockey.

For Dr. Dante Salto, my committee member. Dr. Salto joined my committee after another member was unable to continue serving, something for which I am incredibly grateful. His content expertise improved this dissertation in countless ways. Although we have only known each other for some months, his kindness and thoughtfulness are unmistakable.

For Chris Birr, school psychologist, and valued colleague. Chris was not only instrumental in conducting the study this dissertation was born of but was vital to the completion

of this work. His willingness to consult, brainstorm, and provide access to his incredible network of professionals made this dissertation happen.

For my mother, Jane Reinke who is simply the reason I am here to write these acknowledgments. My mother taught me how to laugh often, love freely, and persevere through hardship. Her ability to make everyone around her feel special is an inspiration. Her example has made me a better person.

For my grandmother, Betty Reinke (Grammy). My love for Grammy is boundless. She has seen me through good times and bad, never wavering. When I feel myself faltering, I think of her smile and the note on the door. I could not have completed graduate school without her. She is also the most loving, gracious, and caring person I know.

For my late grandfather, Gerald Reinke (Grampy). He did not live to see the end of my journey through graduate school, but I know how proud he was. Grampy gave me the greatest gift of my life, his ethical will. My greatest desire is to carry on his legacy.

For my aunt, Jill Reinke (Aunty Jill). Aunty Jill carried me through the hardest moments of my life. She brings joy to those around her and shares her love and talent freely. Aunty Jill gave me The Marx Brothers, guitar, baseball, and Spinal Tap. What else is there to say?

For my aunt Jenny (Jen) and my uncle Greg. Jen and Greg showed me the world and raised me like their own son. I am never safer than with Jen and never laugh harder than with Greg. They inspired in me a love for travel, food, and new experiences. I cannot thank them enough.

For Mark and Mary, my family by love if not blood. Mark and Mary are the rock, the safe landing, the port in the storm. Mary is a mother to me, and Mark a father. They gave me everything a son could ever ask for. I would not be here without them.

For all my teachers and mentors over all these years, in particular Bruce Winchester, Sharon Hostak, Dr. James Rafferty, and Dr. Ryan Schafer. Bruce Winchester taught me to love reading and to think critically. Sharon Hostak believed in me when I did not believe in myself and gave me the opportunity to grow when I needed it most. Dr. James Rafferty inspired in me a love for statistics and research. Dr. Schafer showed me how meaningful the practice of school psychology can be.

For my friends and colleagues in graduate school, especially Amber, Sara, and Rachael. I am so inspired by each one of you. The many hours I have spent with Amber have not only brought me incredible happiness but have made me a better psychologist and a better person. She is loving, intelligent, and fiercely loyal. My life is better with her in it. Sara is brilliant, driven, and the funniest person I know. There may be no better person to grab a beer with, and that is a high compliment. Rachael, through ups and downs, through thick and thin, you were there for me and I will never forget it.

For all my friends outside of graduate school especially, Claire, Jake, Joe, and Nate. Claire is a light in dark places, an example to follow when I am lost; I would not have gotten into graduate school without her. Jake's intellect, his fearlessness, and his irresistible personality are gifts to me and everyone we know. He gave me the happiest trip of my life, right when I needed it most, and I will forever be in his debt. Joe is the most compassionate person I have ever met. His example as a person, a student, and a professional has driven me in ways he will never know. There are not enough late nights to talk, debate, and commiserate with Nate. I treasure his ideas,

insight, and perspective above all else. All of you wake me up to the good stuff I wasn't paying attention to, and I can never thank you enough.

For John, whose friendship is utterly essential to me. John is a friend, a colleague, and an inspiration. He is a brilliant school psychologist who constantly opens my mind to new ideas, a loving husband and father who shows me the man I want to be, and a consummate friend who has given me more love and companionship than any person could ask for. Whether it is Atlanta or Louisville, my couch or his, some of the greatest moments of my life have been with John. He has made the last six years the best of my life.

For Logan, my best friend and brother to me since elementary school. There is no adequate way to acknowledge Logan's influence on my life. He is an educator, author, husband, father, and loving comrade. His intellect is unmatched, his curiosity boundless, and his compassion for others remarkable. He carries me even when we are far apart and, in those moments, I picture myself walking again with him through the fog on the banks of the St. Lawrence, or from bar to bar on Water Street, or to our seats at an Admirals game and I smile. Logan has made my life extraordinary, and I am thankful every day to have a best friend who gives me so much.

Finally, for Madeline. Madeline has been with me since the start of this dissertation. She is, simply, everything to me. From helping me write, supporting me through APPIC interviews, to moving across the country with me, she takes my dreams and makes them possible. She makes the bad times tolerable and the good times phenomenal. She gives my life vibrance and verve and brings me joy and happiness every single day. I am grateful for every second I have with her. She is my favorite person on this earth. Thank you, Madeline, and I love you.

CHAPTER ONE

Introduction

According to the National Center for Education Statistics (NCES), the percentage of American middle-school students performing below the National Assessment of Educational Progress (NAEP) basic standard declined considerably between the years 1990 and 2013 (NCES, 2020). The NAEP classifies students as meeting (a) below basic standards, (b) basic standards, (c) proficient standards, or (d) advanced standards. In 1990, 48% of all middle school students performed below basic standard, while by 2013, that percentage had decreased to 26%. Despite this progress, NCES has tracked an increase in middle school students performing at the below basic NAEP standard since 2013. Since that year, the rate of students achieving at the below basic level has increased from 26% to 31%, with the percentage increasing year over year.

This decrease in the number of students with basic math skills appears to be increasing educational outcome inequality overall. The percentage of students at the proficient or advanced standard has not changed substantively over the past seven years (together, students at these standards accounted for 42% of all middle schoolers in 2013, 41% in 2014). It seems that students exceeding basic standards are continuing to excel, while students meeting basic standards are beginning to fall into the below basic standards category.

An effective approach to support students not meeting NAEP standards is early mathematics intervention (Chodura, Kuhn, & Holling, 2015; Cheung & Slavin, 2011; Burns, Codding, Boice, & Lukito, 2010). The NAEP achievement statistics reported by the NCES are derived from year-end state tests such as the Wisconsin Forward Exam. With few exceptions, all students in public education must take part in year-end achievement testing. This important testing regime helps policy makers, administrators, and even individual educators identify trends,

track progress, and understand how districts perform relative to one another academically. Early intervention for students who are struggling to succeed in mathematics (almost one-third of all middle schoolers) can increase performance on state-wide testing and better prepare students to advance on to more complex mathematical concepts and subjects.

In response to the relatively poor academic achievement of students in the 1990's and a host of other factors, researchers and practitioners began to develop alternative models for implementing interventions in schools. This movement culminated in the passage of the Individuals with Disabilities Education Improvement Act (IDEIA, 2004), which for the first time allowed schools to use Response to Invention (RTI) to provide early intervention to students who were at-risk of academic problems (Fuchs & Fuchs, 2006). The act went further and allowed schools to use up to 15% of the federal funds they receive for special education services on RTI. Although RTI was initially developed to evaluate students with specific learning disabilities, as of the passage of IDEIA, the framework could be expanded to address a host of academic problems. The RTI framework was itself was then blended with Positive Behavioral Interventions in Schools to form the comprehensive Multi-Tiered Systems of Support (MTSS) model (McIntosh & Goodman, 2016; Stoiber & Gettinger, 2016).

The MTSS model is a framework that helps schools and districts identify at-risk students and provide appropriate social-emotional and academic evidence-based interventions as needed. In a graduated fashion, at-risk students are provided tier one, two, or three interventions, depending on the severity of the problem. If students do not respond to these interventions, they are typically moved up to a more intensive tier. While the process of conducting MTSS on paper is relatively straightforward, in practicality, there are a multitude of considerations that affect the utility of the framework. Early identification of students with difficulties is one aspect of the

MTSS framework that has required substantial effort to implement in recent years (Glover & Albers, 2007).

Identification of academically at-risk students is a holistic effort. Classroom teachers, parents, or even students themselves may identify areas for growth that should be addressed in the MTSS framework. Formative and summative classroom test scores, patterns of behavior, and even symptoms of decreasing mental health may be early signs of academic problems that may warrant a referral for intervention (Grimm, 2007). The most researched and validated method of identifying at-risk students is evidence-based academic screening measures. The developers of validated screening measures purport that they have strong psychometric properties, provide adequate diagnostic accuracy, and are easy to administer.

As most screening measures are provided by for-profit companies, it can be difficult for researchers to take reported psychometric properties and diagnostic accuracy indices at face value. Indeed, recent research (which this dissertation was developed from) by Klingbeil and colleagues (Klingbeil et al., 2019) found that without creating local cut scores, two of three screening methods studied did not provide even minimum levels of diagnostic accuracy in some middle school grades.

One of the screening methods that underperformed in terms of diagnostic accuracy was Curriculum-Based Measures (CBM) provided by Pearson. VanDerHeyden and colleagues (2017) have suggested that the relatively low diagnostic accuracy of CBM is offset by its usability (screening often takes under three minutes) and relative lack of expense. The question of how to weigh diagnostic accuracy and expense is one perfectly suited to cost-effectiveness analysis (CEA). Put simply, CEA is a method for quantifying both the effectiveness and cost of an intervention, program, or screening measure in a single metric. In this fashion, researchers can

identify the approximate cost of obtaining one additional unit of effect. This statistic (called an incremental cost-effectiveness ratio) can be used to directly compare separate screening measures—for example, we can see how a cheap but relatively inaccurate screening measure compares to an expensive but very accurate one.

Statement of the Problem

Researchers have understood the importance of cost-effectiveness analysis in education for some time, although empirical work on this topic is rare (Levin & McEwan, 2001). To conduct high-quality CEA, researchers need to both know the efficacy and cost of a program, intervention, or screening method. Determining the efficacy (i.e., diagnostic accuracy) of a screening method requires school psychologists to assess the number of true positives, false positives, true negatives, and false negatives produced by a screening measure (Christ, Nelson, & Van Norman, 2014). These metrics are produced by comparing the predicted outcomes of a screening measure compared with actual student performance on a criterion assessment (typically a year-end state-wide test).

If a screening measure predicts a student will meet proficiency standards and that student demonstrates proficiency on a criterion test, the screening measure is said to have produced a true negative (negative meaning that the student is not at-risk). If a student is predicted to be at-risk for performing below proficiency standards and that student does perform below proficiency standards on the criterion assessment, the student is classified as a true positive. The inverse of these screening decisions is used to determine false positives and false negatives. Diagnostic accuracy can also be gleaned from prior research and published reports, although Klingbeil and colleagues (2019) demonstrated that there can be wide differences between reported diagnostic accuracy and actual diagnostic accuracy of screening measures when administered in an applied

setting. As such, it is best practice for researchers and school psychologists to assess screening efficacy at the local level.

While there has been significant research on the efficacy of screening methods, there is relatively little on costs. Prior surveys of school psychologists have found that the most reported obstacle to the implementation of a new intervention or program is a lack of financial resources (Forman, Olin, Hoagwood, Crowe, & Saka, 2009). Forman and colleagues (2009) also identified the second-largest hindrance to implementation as a lack of school psychologist time. These obstacles are rarely accounted for in efficacy research or are only vaguely addressed (Levin, 2001). The same is true for studies examining universal screening methods (VanDerHeyden & Burns, 2018). Although researchers often identify cost as an essential consideration when selecting screening methods (Glover & Albers, 2007), actual inputs of time, money, and equipment have not been yet quantified. By conducting a CEA, school psychologists can accurately measure the costs, both apparent and hidden, of universal screening methods while simultaneously emphasizing diagnostic accuracy.

Taken together, the need for evaluating efficacy at the local level and accurately assessing costs makes performing high-quality CEAs on academic screening measures difficult. This challenge has led many researchers studying academic screening to attempt to evaluate costs (Klingbeil et al, 2017), but largely not in granular detail. Other researchers have attempted to conduct rigorous CEA in mathematics, but on interventions, not screening methods (Barrett & VanDerHeyden, 2020). To my knowledge, no peer-reviewed study has yet to conduct a rigorous CEA on mathematics screening methods.

Overview of the Study

The purpose of this study is to demonstrate that CEA can be used to help discriminate among screening measures that tend to have relatively similar levels of diagnostic accuracy. By finding screening methods that are both effective and efficient, schools and districts can make better use of limited resources and help students achieve. Some studies have attempted to quantify the costs of screening but often overlook the many ingredients required to conduct a CEA. This is the first study that conducts a CEA on universal mathematics screening approaches in middle school.

This study uses both novel and extant data from prior research. To determine the diagnostic accuracy of three individual screening measures and two combinations of measures, extent data were used. My colleagues and I collected data from a suburban school district in southwest Wisconsin for prior research (Klingbeil et al., 2019). A total of 1,587 middle school students participated in the study, contributing screening data from the Wisconsin Forward Exam, Measures of Academic Progress, and Math Concepts and Applications curriculum-based measures. Multiple imputation was used to address missingness. Sensitivity, specificity, and other metrics of diagnostic accuracy were calculated, and the researchers built two multiple linear regression models to examine combinations of screening methods. This study used these data to calculate diagnostic odds ratios, a measure of screening accuracy that was appropriate for use in CEA. This study also quantified the cost per student of 27 unique ingredients needed to administer the three screening measures evaluated to compute incremental cost-effectiveness ratios. These ratios were then used to make direct comparisons of screening approaches in terms of cost-effectiveness. In addition to these primary analyses, I conducted two sensitivity analyses to test the robustness of the findings, calculated coefficients of variation to evaluate CEA's

ability to increase discrimination among measures, and measured relative changes in costs and diagnostic accuracy as screening methods were combined.

Potential Contribution to the Research

There are multiple screening options available to schools and districts with relatively small differences in diagnostic accuracy, the debate over how to select the most accurate and efficient screening process continues (Albers, Glover, & Kratochwill, 2007; Hummel-Rossi & Ashdown, 2002; Gersten et al., 2012). This study may help remedy this debate by evaluating whether CEA could provide additional relevant information for decision makers. By conducting a CEA on universal mathematics screening approaches in middle school, this study may provide a proof of concept for the technique, demonstrate that additional useful information that can be gained beyond metrics of diagnostic accuracy, and guide schools and districts as they attempt to find the best screening approach for their students.

CHAPTER TWO

Literature Review

This study examines the cost-effectiveness of five different screening approaches involving three unique screening methods. This literature review begins with an introduction of universal academic screening, explains various metrics of diagnostic accuracy, and provides an overview of the debate over cost and efficacy in the universal screening literature. Next, this literature review will provide an overview of cost-effectiveness analysis (CEA), the incremental cost-effectiveness ratio, and applications of CEA in education and special considerations in medication. The final section of this literature review focuses on the potential for applying CEA to universal mathematics screening in middle school.

Universal Academic Screening

Universal academic screening is a crucial component of the multi-tiered systems of support (MTSS) framework (Glover & Albers, 2007). Accurate, efficient identification of students at-risk for academic failure allows school psychologists to implement interventions (Elliott, Huai, & Roach, 2007). Early identification of academic problems has long been known to improve student outcomes both in the long and short term (Walker & Shinn, 2002). In the past, school psychologists practiced a "wait-to-fail" model, whereby students would receive intervention only after failing a class or criterion measure (Albers, Glover, & Kratochwill, 2007). This model of service delivery can have deleterious effects on student outcomes and does not provide the conditions needed for academic success. Albers and colleagues (2007) also commented that identifying different levels of risk (i.e., current functioning, increasing or decreasing difficulties, etc.) is essential to successful universal screening, suggesting that high-quality universal screening may need to be conducted at multiple time points to monitor

progress. As such, schools have moved towards a universal screening model, where student progress is monitored throughout the school year.

Although increased universal academic screening does appear to result in better outcomes for students, Glover, and Albers (2007) note that an essential component for successful screening is usability. The authors note that, while a particular screening measure may have adequate diagnostic accuracy, it may not be feasible to administer. Importantly, Glover and Albers (2007) first facet of usability is cost. Conducting high-quality universal academic screening with minimal resources continues to be a challenge. Researchers and practitioners have searched for ways to decrease the time and resources required to conduct screening, with methods such as curriculum-based measurement (CBM), computerized adaptive testing (CAT), and even single item rating scales proposed (Stormont, Herman, Reinke, King, & Owens, 2015; Salinger, 2016). Despite this work, the search continues for new approaches to screening and for new metrics that may discriminate among the ever-growing number of universal academic screening methods.

Screening in Mathematics. Although many schools conduct universal screening in multiple areas (e.g., reading, social-emotional skills), the focus of this study is on screening to predict math risk. Universal screening in mathematics has drawn significant consideration from scholars and school psychologists alike (Gersten et al., 2012). Various researchers have suggested different approaches to mathematics screening, although most have agreed that the underlying skill measured is number sense (Gersten et al., 2012). Number sense—as defined by Okamoto and Case (1996), is the process whereby students gradually gain a more complex and nuanced understanding of numbers and their manipulation. Despite some consensus on the underlying skill measured, scholars have debated how best to evaluate it. Some researchers advocate for specific, skill-based measures that gauge only one aspect of number sense at a time

(e.g. two-digit by two-digit division, the order of operations, etc.). Others have suggested that mathematics screeners should be broader and more similar to criterion measures. Fuchs, Fuchs, and Zumeta (2008) have argued that mathematics screeners should consist of a variety of grade-level problems representing the core mathematics standards that students are expected to meet.

In any case, research into the most effective and efficient mathematics screening measures has continued into the present day. Klingbeil and colleagues (2019) conducted one of the more recent studies of universal screening methods in mathematics for middle school students. VanDerHeyden, Codding, and Martin (2017) also conducted a study of mathematics screening methods in the elementary grades. Scholars continue to attempt to find new screening methods and data analytic procedures to decrease the number of students who are incorrectly classified during universal screening.

Diagnostic Accuracy. The most essential consideration for selecting screening methods and measures is the diagnostic accuracy of resulting predictions about student risk (Johnson et al., 2007). Researchers often use indices such as false positives, false negatives, sensitivity, specificity, and positive and negative predictive values to evaluate screening methods (Christ, Nelson, & Van Norman, 2014). Universal screening measure results determine if a student is at-risk vs. not-at-risk, with four outcomes possible. The first two outcomes, a true positive (TP) and a true negative (TN) occur when a screening method accurately categorizes at-risk and not-at-risk students. A TP occurs when a student is found to be at-risk on a screening measure and subsequently fails a criterion measure. A TN occurs when a student is categorized as not-at-risk on a screening measure and goes on to pass a criterion measure. Both TP and TN represent correct screening decisions, and a screening measure with high diagnostic accuracy will result in the vast majority of screening decisions being either TP or TN.

A criterion or gold standard measure is essential for research on the efficacy of screening approaches. Without such a measure, there is no method for determining how well a screener categorizes students. Despite this fact, there is no current widely accepted criterion measure used nationally. Individual States develop their own end-of-year achievement tests, often based on their individual State standards. As such, academic screening researchers often use these statewide tests as criterion or gold standard measures. These end-of-year Statewide tests, mandated by federal legislation and used to determine whether a student is at grade level, suffice as criterion measures in the absence of a broader nationwide standard.

False positives (FP) and false negatives (FN) comprise incorrect screening decisions. A FP occurs when a student is identified as at-risk during screening, but who then passes a criterion measure. Too many FP leads to schools expending valuable resources in providing interventions to students who likely do not need them. The inverse of this is a FN, whereby a student is categorized as not-at-risk on a screening measure, but who then goes on to fail a criterion measure. A FN is often considered the costliest screening error as it could prevent a student who likely needed early intervention from receiving it because the student was expected (incorrectly) to pass a criterion measure (Glover & Albers, 2007).

Sensitivity (SE), specificity (SP), positive predictive values (PPV), and negative predictive values (NPV) are all metrics of diagnostic accuracy based on the relative number of TP, TN, FP, and FN a screening measure produces. Each of these indices is on a scale of 0 to 1, with values closer to 1 representing higher diagnostic accuracy. Sensitivity is the relative ratio of students who are correctly identified as at-risk out of all students who fail a criterion measure. Sensitivity is calculated by dividing the number of TP by the sum of TP plus FN, which together represent all at-risk students in a sample (TP / [TP + FN]). Specificity is the relative ratio of

students who are correctly identified as not-at-risk out of all students who pass a criterion measure. Specificity is calculated by dividing the number of TN by the number of TN plus FP, which together represent all not-at-risk students in a sample (TN / [TN + FP]). Researchers often consider sensitivity the more important measure of diagnostic accuracy due to the potential harm associated with not intervening on students who require it. As such, acceptable sensitivity values have a higher threshold than acceptable specificity values (.90 for SP and .70 for SE; Johnson, Jenkins, Petscher, & Catts, 2009; Kilgus, Methe, Maggin & Tomasula, 2014).

Positive predictive values represent the ratio of students who were identified as at-risk on a screening measure to those who were truly at-risk (as measured by failing a criterion measure). Positive predictive values are calculated by dividing the number of TP by the number of TP plus FP, which together represent all students who were identified as at-risk in a sample (TP / [TP + FP]). Finally, negative predictive values represent the ratio of students who were identified as not-at-risk to those who were truly not at risk (i.e., they passed the criterion measure). Negative predictive values are calculated by dividing the number of TN by the number of TN plus FN, which together represent all students who were identified as not-at-risk in a sample. As with sensitivity and specificity, positive predictive values are generally considered to be more important in screening than negative predictive values.

Screening Measures and Methods. As school psychologists continue to understand the importance of universal mathematics screening in schools, researchers have attempted to provide an increasing array of screening methods from which to choose. Curriculum-based measures (CBMs) assess discrete skills that students encounter in the classroom (VanDerHeyden et al., 2017) but bear little resemblance to end-of-year criterion measures, particularly in Grades 3 through 8. Computer adaptive testing (CAT) more closely resembles criterion measures such as

statewide achievement tests. However, some have noted that CAT may be too resource-intensive for universal screening (January & Ardoin, 2015). Others have advocated for using the previous year's criterion measure as a screening measure prediction performance on the following year's criterion measure (Fuchs, Fuchs, & Compton, 2010). All approaches have strengths and weaknesses, further complicating the issue of method selection.

Curriculum-based measurement. Curriculum-based measures have been widely used for universal screening over the last decade (Ball & Christ, 2012; Deno, 2003; Prewett et al., 2012). Originally designed for progress monitoring, a meta-analysis of CBMs indicates that this method does have adequate evidence of predictive validity ($r \geq .68$; Yeo, 2010). CBMs have also been shown in some studies that have adequate diagnostic accuracy for high-stakes decision-making (VanDerHeyden et al., 2017). Both the meta-analysis by Yeo (2010) and the study by VanDerHeyden and colleagues (2017) both focused on reading, not mathematics; in general, more research has been conducted on reading CBM than math CBM. However, other research has suggested that CBMs may not be discriminative enough to be an effective screening measure (Klingbeil et al., in press; Shapiro & Gebhardt, 2012). While desperate results on CBM have been found in the literature, there is relatively little research on the subject (especially at higher grades and in mathematics) making a direct comparison of study results difficult. For example, although VanDerHeyden et al. (2017) and Klingbeil et al. (2019) came to different conclusions regarding CBM, the studies in question did not have similar samples in terms of age/grade range making any generalizing difficult. Despite this, advocates for CBM argue that practitioners can administer probes at a more frequent rate than other screeners, are much less time and resource-intensive than other screening methods such as computerized adaptive testing, and allow educators to hone in on specific math deficits that can be seen in CBM that may not be

uncovered using other methods. While CBM is almost certainly less costly in terms of administration time (mathematics CBM probes can often be administered in under 15 minutes), researchers have not examined other costs such as training and scoring.

***Computer adaptive testing*.** Computer adaptive testing (CAT) is a newer method of screening that has been widely adopted in schools (Cope & Kalantzis, 2016). CAT is unique in that all students in a classroom or school will not receive the same screening questions. When a student answers a CAT item correctly, they will next be shown a more difficult item; if a student incorrectly answers a CAT item, they will next be shown an easier item. In this way, CAT allows for a wider range of scores, with fewer floor and ceiling effects than traditional CBM due to the individualization of each test (Van Norman, Nelson, & Parker, 2017). Despite these advances, CAT appears to be much more resource-intensive on its face than CBM, leading some to suggest that it may not be the best screening method school psychologists can select (VanDerHeyden & Burns, 2018). Examined for technical adequacy alone, CAT has been shown to provide acceptable diagnostic accuracy for high-stakes decisions (Shapiro & Gebhardt, 2012).

***Prior state test performance*.** One method of universal academic screening that is attracting interest is the previous year's criterion or summative evaluation. Beginning with the No Child Left Behind Act of 2001, the federal government required states to conduct year-end summative evaluations to demonstrate Adequate Yearly Progress (AYP). Practitioners can use these criterion measures (i.e., what universal screening measures attempt to predict) themselves to predict future academic risk (e.g., Vaughn & Fletcher, 2012; Van Norman, Nelson, & Klingbeil, 2017). Indeed, Van Norman and colleagues demonstrated that, for middle school-aged students (Grades 4 through 8), previous year state test scores have approximately the same diagnostic accuracy in the area of reading as the Measures of Academic Progress (MAP), a CAT

screening method. Klingbeil and colleagues (in press) found similar results in mathematics, where previous year state testing was the single best screening method for middle school students after creating local cut-scores.

Local cut-scores. Another method for conducting universal academic screening is the creation of local cut-scores. It is important to note that this method is only useful after screening data has been collected. CBM, CAT, or other screening methods directly measure student performance, while the creation of local cut-scores allows practitioners to adjust the SE or SP of these screening tools based on previously collected data. To create a locally derived cut-score, practitioners must have data from both a screening measure and the criterion measure. Using receiver operator characteristic (ROC) analysis, practitioners can adjust either the SE or SP of a screening measure at the expense of the other. Creating local cut-scores cannot increase both SE and SP simultaneously. As false negatives are considered the most egregious screening error, most practitioners choose to increase the SE of a measure at the expense of SP. Researchers have shown that creating local cut-scores can raise the SE of a screener while still maintaining acceptable SP for CAT (Van Norman, Nelson, & Klingbeil, 2017; Klingbeil et al., 2019), CBM (Straight, Smith, & McQuillin, 2018), and even previous year state testing (Nelson, Van Norman, & Lackner, 2016).

Multiple Screening Measures. Within an RTI framework, schools follow what has been called the direct route (DR) for intervention (Johnson, Jenkins, & Petscher, 2010). Using the DR approach, schools place students found to be at-risk on a single screening measure into a Tier 2 intervention. Because DR screening relies on a single screening measure to classify students, these individual measures must be highly accurate. Other screening models are more robust to less accurate screening measures. For example, schools may choose a progress monitoring (PM)

model where students who are identified as at-risk are closely monitored over time to determine if they are progressing, regressing, or remaining stationary. The PM model is more robust than the DR model against inaccurate screening measures as it allows students who may have been incorrectly identified to demonstrate progress.

Another screening method that has been proposed as more robust than the DR approach is the multiple measures (MM) model (Gersten et al., 2009). The MM approach combined multiple screening measures using regression. The linear combination of measures is thought to be more robust to inaccurate screening measures as there are multiple sources of screening data that the models account for. Johnson, Jenkins, and Petscher (2010) demonstrated that combining screening measures resulted in a 2% increase in classification accuracy in the domain of reading. Klingbeil et al., 2019) replicated these findings for mathematics and showed a small increase in area under the curve (AUC) when multiple measures were combined. Despite these results, not all researchers agree that combining screening measures results in better screening decisions. VanDerHeyden (2013) noted that student scores on different screening measures tend to be very highly correlated and therefore additional screening measures will likely not improve diagnostic accuracy significantly. Any small increase in diagnostic accuracy may not warrant the cost of administering more than one screening measure. Indeed, VanDerHeyden (2011) has argued that schools collect too much universal screening on the assumption that classification errors will decrease, a premise that is incorrect at least some of the time.

Usability of Screening Methods. Although most practitioners now recognize the importance of universal academic screening, selecting an appropriate screening method remains a problem (Glover & Albers, 2007). Glover and Albers (2007) published a seminal paper laying out the conditions that accurate and efficient screening methods should meet. The authors

categorized these considerations into three domains, (a) appropriateness for intended use, (b) technical adequacy, and (c) usability. A good deal of research has attempted to discover which screening methods have the highest technical adequacy (VanDerHeyden, Codding, Martin, 2017; Klingbeil, Nelson, Van Norman, & Birr, 2018; Klingbeil et al., in press). Additional research has investigated which screening methods are most appropriate in each situation (VanDerHeyden, 2013). But very little consideration has been given to how to select screening methods based on usability (VanDerHeyden, Burns, & Bonifay, 2018).

Glover and Albers (2007) describe the considerations that fall under the domain of usability as (a) a balance of costs and benefits, (b) feasibility of administration, (c) acceptability (d) infrastructure requirements, (e) accommodation needs, and (f) utility of outcomes. Although the considerations of acceptability and utility of outcomes almost certainly fall into the category of social validity (Wolf, 1976), researchers could address Glover and Alber's (2007) other considerations using methods such as cost-effectiveness analysis (CEA). Many scholars have attempted to address the issue of costs when conducting universal academic screening (Klingbeil et al., 2017; VanDerHeyden et al., 2017; VanDerHeyden et al., 2018). One example can be seen in a study by VanDerHeyden and colleagues. The authors attempted to compare both decision accuracy and screening costs for multiple mathematics screening methods in elementary school. VanDerHeyden et al. (2017) suggested that the most cost-effective screening measure was the previous year's criterion measure as it required no additional administration time to collect data. Despite this, schools often use screening measures for multiple purposes, such as tracking within-year growth. The previous year's state test cannot be used for these purposes and so it is important to consider the cost-effectiveness of all screening methods available. And although presuming that the previous year's state test will be the most cost-effective is a completely

defensible hypothesis, costs were not systematically collected or analyzed in this study and definitive statements likely cannot be made as of yet.

Another example can be found from Klingbeil and colleagues (2017). The authors calculated the amount of instructional time (in minutes) saved by not conducting running records on top of other screening measures. The amount of instructional time saved (between 270 and 540 minutes) was substantial and an important metric of screening cost. However, up to now, most scholarly writing on the costs of universal screening methods has focused on broad estimates of the time required to collect assessment data, and not included a systematic approach to evaluating costs.

Instructional time is only one method of defining costs. The costs of conducting universal screening could be more comprehensively measured by carefully considering all of the myriad inputs required to administer screeners effectively. Universal academic screening certainly requires instructional time to complete but may also include the time required for scoring and interpreting, the time needed to conduct training or in-service, materials such as computers or paper, facilities such as classrooms, and other inputs such as license fees. Indeed, lost classroom instructional time may only be one small fraction of the cost of screening. For school psychologists to make informed decisions, practitioners should take into account both the effectiveness and the true costs of universal screening measures.

Cost-Effectiveness Analysis

Practitioners may confuse cost-effectiveness analysis (CEA) with cost-benefit analysis (CBA), although both can be utilized in the school setting. Cost-benefit analysis differs from CEA in that benefits are measured only monetarily. For example, a CBA on interventions preventing high school dropout would quantify the benefit of an intervention in terms of cost

savings to society; people who graduate high school generally use fewer governmental resources (e.g., food assistance, subsidized housing, etc.) than those who do not (Levin & Rouse, 2012). In this fashion, CBA allows researchers to compare the amount of money required to implement an intervention with the amount of money generated or saved by the successful completion of said intervention. If the benefits (monetary) outweigh the costs, the intervention or program may be adopted. Cost-effectiveness analysis uses effect size metrics as outcome measures, not monetary benefit.

Cost-effectiveness analyses first began as a method for selecting weapons systems during the height of the cold war (Levin, 2001) and were quickly adopted in healthcare. Despite the widespread use of CEA in these and other fields, researchers in education have been slow to embrace the method. Indeed, Levin (1991) conducted a survey that found that less than 1% of presentations at a national education conference implemented CEA over three years. Since that time, the use of CEA in education has grown substantially, with the number of peer-reviewed education articles discussing CEA increasing year over year (Levin, 2001). Despite this welcome change, to this point, CEA has not been used in the study of universal academic screening methods.

Cost-effectiveness analysis may be a powerful tool that has been underutilized by those studying universal academic screening. Conducting a CEA allows a practitioner to combine measures of effect with metrics of cost so that alternatives in intervention, curriculum, programming, etc. can be directly rank-ordered. This rank ordering has helped researchers determine the most cost-effective programs for decreasing high school drop-out, the relative efficiency of interventions at different grade levels, and even the effect of potable drinking water on academic outcomes in underdeveloped nations (Levin & McEwan, 2001). It may be that these

techniques may also help researchers rank order universal screening methods, a task which as of yet has not been undertaken.

The rationale for conducting a CEA is to quantify the ratio of intervention cost divided by intervention effect—researchers refer to this ratio as the cost-effectiveness ratio (ICER). In short, a CEA informs a practitioner of how much it costs to obtain a unit of effect size. Interventions that are expensive and have relatively small effect sizes will have very large ICERs; interventions that are less expensive, but which have relatively large effect sizes will have small ICERs. In this way, practitioners can perform CEA on two interventions to identify which has the lower ICER—in other words, which is more cost-effective.

Incremental Cost-Effectiveness Ratios. An ICER is calculated by dividing the total cost of an intervention by the effect size of said intervention. The practitioner inputs the cost per student of a selected intervention and the intervention's effect size (e.g., an effect size reported in peer-reviewed literature or derived from an efficacy study) to obtain a result.

When comparing universal screening methods, it is helpful to understand how the ICER changes as a function of cost and effect size. If two methods have similar costs—if the value in the numerator remains constant—researchers can model ICERs as an exponential function as seen in Figure 1. In practical terms, this means that the ICERs of two screening methods with relatively small effect sizes will be much further apart than those of two screening methods with relatively large effect sizes. As an example, consider two universal screening measures that cost \$100 per student to implement, with the effects of each measured by the diagnostic accuracy metric Diagnostic Odds Ratio (DOR), and with a DOR difference of $DOR_{\text{difference}} = 1.00$. If these two screening measures had relatively small DOR of $DOR = 9.00$ and $DOR = 10.00$, their respective ICERs would be $ICER = \$11.11$ and $ICER = \$10.00$, a difference of \$1.11 per student

to obtain a one unit increase in DOR. However, if these same measures had relatively large DOR (while maintaining a DOR difference of $DOR_{\text{difference}} = 1.00$) of $DOR = 39.00$ and $DOR = 40.00$, their respective ICERs would be $ICER = \$2.56$ and $ICER = \$2.50$, a difference of only \$0.06 per student to obtain a one unit increase in DOR. Other considerations for the application of CEA to universal academic screening are discussed in Chapter 3.

Figure 1. Cost-Effectiveness Ratio as a Function of Effect Size

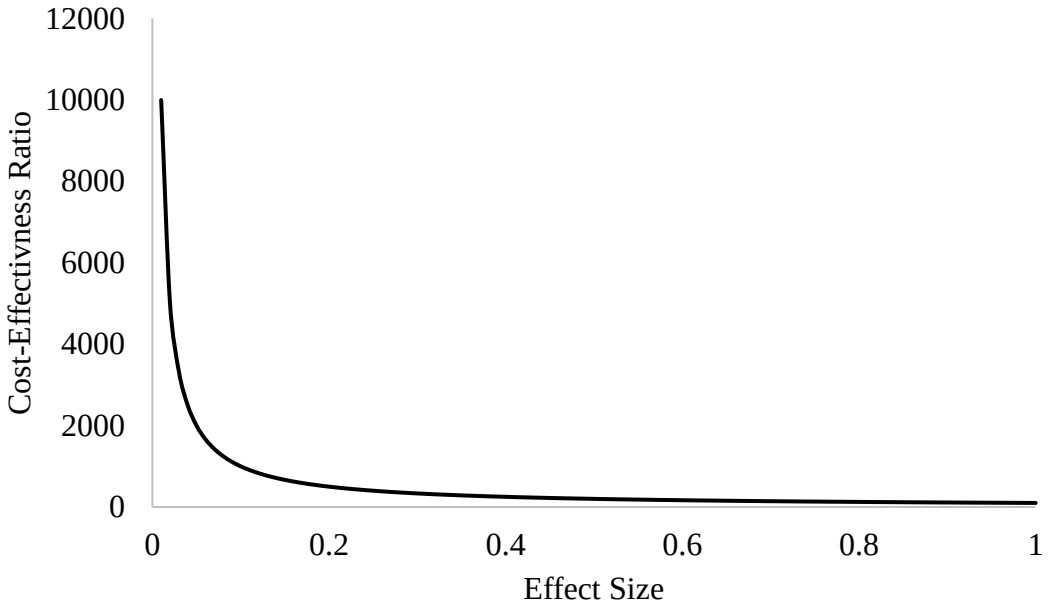

Figure Notes: Cost-effectiveness ratios as a function of effect size under the condition that cost is held constant.

This distinction is important as it guides school psychologists in their decision-making process in selecting from various universal screening methods. There is a strong diminishing return in finding screening methods that are more cost-effective when existing methods already have large effect sizes and if costs remain constant. Many screening methods studied fit both of these criteria as most screening studies have not attempted to quantify costs (Gersten et al., 2012). If costs remain undefined, they are unintentionally held constant as there is no variability. By holding costs constant by not defining them, practitioners are forced to compare screening methods solely using diagnostic accuracy, a practice which may be difficult as many screening measures perform relatively well and there is little variability that can be leveraged to help select a screening method.

When using the ICER to compare two interventions, practitioners must standardize both costs and effect sizes. For example, researchers cannot compare an intervention with an effect reported in terms of r^2 to an intervention with an effect reported in terms of Cohen's *d* until they transpose these values onto the same scale of measure. Once practitioners standardize both costs and effects, ICERs are calculated and used to select the most cost-effective approach.

Determining how to select the most cost-effective approach has been addressed primarily in the medical literature. Cohen and Reynolds (2008) described the cost-effectiveness plane approach. This method involves separating interventions into four quadrants. Two of the quadrants are "Dominant" and "Dominated", reserved for interventions that are more effective and less costly or less effective and more costly, respectively (interventions that are cheaper and better are considered to be "Dominant", the opposite being true for "Dominated" interventions). The two remaining quadrants are for interventions that are more cost-effective overall due to

being less effective but much cheaper, or being more expressive but much more effective. All interventions tested using CEA will fall into one of these quadrants.

Although this is a useful way of conceptualizing the results of CEA, the CE Plane does little to guide researchers in understanding how to select interventions based on ICERs. Paulden (2020), identified four steps to help solve this issue. First, the researcher selects the single most cost-effective intervention (i.e., the intervention with the lowest ICER). Second, the researcher rank orders different interventions by their ICERs from lowest to highest. Third, the researcher evaluates the magnitude of difference between ranked ICERs. Finally, sensitivity or scenario analyses are used to test the robustness of an intervention's ICER. For this study, tasks one, two, and four were used to determine cost-effectiveness. Levin and McEwan (2001) argued that there may be some exceptions to this approach—for instance, one intervention may have a higher ICER, but a practitioner may still select it due to a more robust evidence base—but these exceptions are outside the scope of this paper.

Screening Costs. While it is relatively easy for practitioners to determine the effect size (i.e., diagnostic accuracy) of a screening method by examining relevant literature, it is much more difficult to determine the costs of the same method. Universal screening may have upfront costs for purchasing manuals, training programs, etc. which are easy to identify. Other upfront costs are also identified without much difficulty. For instance, if a screener requires a school to purchase technology such as tablets, a practitioner would very likely recognize the price of the tablets as part of the cost of conducting universal screening. However, other resources such as building or room usage, faculty time investment, materials such as paper, etc. are very often overlooked (Levin, 1975). Researchers refer to these—often hidden—costs as opportunity costs (Levin & Belfield, 2015).

Opportunity costs refer to the costs of not using limited resources for something other than universal screening. An easily understood example of an opportunity cost is teacher time. Imagine a screening method that requires 30 minutes of teacher time to administer. That 30 minutes of teacher time spent conducting the screening represent an opportunity cost; if the intervention was not implemented, it would be expected that the teacher would have 30 more minutes per week with which to consult with peers, receive training, plan curriculum, etc. As teachers have only a limited number of minutes in a workday, any action that takes up time must be taken at the expense of some other action that must now be foregone. A classroom used for ACT preparation can no longer be used for a study hall. A computer used for web browsing can no longer be used for math intervention.

Fortunately, researchers established methods for calculating costs some time ago (Levin & McEwan, 2001). Researchers refer to one such technique as the ingredients method. Using this method, practitioners list every ingredient required to successfully implement an intervention. To accurately calculate the costs of the ingredients (and, by extension, the intervention as a whole), school psychologists could use CostOut, a freely available toolkit created by the Center for Benefit-Cost Studies of Education (Hollands et al., 2015). CostOut allows the user to estimate the cost of a wide variety of ingredients that practitioners require to implement an intervention. A researcher can customize the costs of intervention ingredients in the toolkit for many unique contexts. For instance, the cost of one hour of paraprofessional support varies widely by where in the nation the intervention is taking place. The user can specify that the intervention is taking place in the Pacific Northwest and the CostOut toolkit will update intervention ingredients cost to take geographical variation in paraprofessional salary into account. For example, the average starting salary for a teacher in the District of Colombia is $55,209, compared to $31,418 for a

teacher in Montana (National Education Association, 2018). If researchers conducted a CEA using the national average starting teacher salary ($39,249) instead of a state or regional average, the cost of teacher time could be under- or over-estimated by 41% or 20% respectively.

This summary of CEA is cursory and does not include the wide variety of techniques, alternatives, and supplementary analyses currently in the literature. Although the scope of this dissertation is limited to basic CEA, the methods described are only the most rudimentary available to school psychologists. Despite this, even the most basic CEAs are not regularly conducted in educational settings (Levin & Belfield, 2015; Hunter, DiPerna, Hart, & Crowley, 2018).

Cost-Effectiveness in Medicine

Educational researchers should be aware of an unintuitive finding from the medical literature on CEA that may have substantial impacts on education. Indeed, researchers in the medical field have long been enthusiastic about the applications of CEA and have provided much of the new methods and techniques that are subsequently adopted in education. Surprisingly, long-term studies of medical innovations—especially new drugs and technologies—have shown that using CEA increases healthcare spending (Mitchell, 2002). Research has shown that new drugs and technologies are usually only adopted when they have been found to have lower incremental cost-effectiveness ratios (ICER) than existing drugs and technologies. However, in one literature review, of all new medical interventions adopted due to having lower ICERs, only approximately 2% of these had lower ICERs because they were cheaper than existing interventions while maintaining or increasing effectiveness. In other words, medical professionals adopt most new interventions because they are more effective (leading to better ICERs) but are almost always more expensive as well (Arbel & Greenberg, 2016). Because of

how ICERs are calculated, researchers consider a new intervention to be more cost-effective—even if it is more expensive—as long as effectiveness also increases dramatically. The best outcome in medicine and education would be the adoption of new interventions that cost the same or less as current interventions while maintaining or increasing effectiveness; the medical literature suggests that this rarely happens. And although medical professionals may consider these new interventions to be more cost-effective based on the ICERs, they are almost always more expensive than what they are replacing. Because of this, the CEA approach in medicine has been criticized as unsustainable as costs will increase indefinitely as long as effectiveness does as well.

It is unclear at this point whether the same pattern would emerge when using CEA in education. Medical researchers have been using CEA rigorously for at least the past three decades, whereas the literature on CEA in education is in its infancy. At this point, it appears logical to assume that education researchers will be able to conduct CEA on existing interventions to make intervention selection more efficient without increasing costs. Additionally, the costs of school-based interventions are orders of magnitude cheaper than medical technologies and drug development, and education researchers may never encounter the issue that medical researchers have. However, it is certainly possible that there may come a time when researchers develop increasingly effective but increasingly costly school-based interventions that cause costs to spiral upwards as in medicine. More research is needed in this area to fully understand the complex interplay between intervention costs and effects in education to answer these questions. At this point, medical researchers suggest that it is best to talk about CEA, not in terms of cost-saving, but in terms of obtaining greater overall value (Neumann, 2004). This language may also be useful in education research.

One study conducted in the medical literature closely resembles this proposed dissertation. Keren and colleagues (2002) conducted a CEA on statewide universal hearing screening for newborns. This study provides an important proof of concept for the application of CEA to universal screening, even though it is unrelated to education. The study compared universal screening at birth with passive screening at six-month—researchers consider passive screening to be when caregivers seek out hearing screening after noticing hearing issues. The authors found that the ICER of passive screening ($69,000) was much higher than the ICER of universal screening ($44,000). Although the cost of passive screening is much less to society, the detrimental effects of no early intervention resulted in a higher ICER than universal screening.

Cost-Effectiveness in Education

Recently, researchers in education and school psychology have shown a renewed interest in the study of cost-effectiveness. Barrett and VanDerHeyden (2020) recently published a study on the cost-effectiveness of a class-wide mathematics fluency intervention. The 15-minute intervention included components such as peer coaching, independent practice, immediate feedback, and a class-wide contingency based on performance. The purpose of the study was not to compare the ICERs of multiple interventions, but instead to get an accurate understanding of the costs associated with implementing the intervention. Importantly, the authors did examine how the intervention ICERs changed as a function of race, sex, socioeconomic status, special education status, and educational risk level. Barrett and VanDerHeyden (2020), writing on the implications of this study, noted that their research provided data for stakeholders to help select among mathematics interventions, a very similar approach that is taken in this dissertation.

Hunter and colleagues (2018) conducted another recent study of CEA in education. The researchers examined the Social Skills Improvement System—Classwide Intervention Program

(SSIS-CIP), a universal social-emotional intervention. Interestingly, the authors did not compare the SSIS-CIP to another intervention, instead opting to examine the cost-effectiveness of a single intervention across grades. Using the ingredients method, the authors estimated the cost-effectiveness ratio (ICER) for implementing the SSIS-CIP in both Grade 1 (N = 60) and Grade 2 (N = 38) classrooms. The authors found that the average cost of implementing the SSIS-CIP across grades (and including start-up and maintenance costs) was $18.99 per student. However, the ICER for Grade 1 students was $105.50 compared to an ICER of only $52.75 for Grade 2 students. Put another way, when implementing the SSIS-CIP intervention, it cost nearly twice as much money to obtain a unit of effect in Grade 1 as it did in the second grade. Based on the findings that the SSIS-CIP was much more cost-effective in Grade 2, practitioners may decide to choose a different intervention for students in Kindergarten or Grade 1.

In earlier research, Hollands et al. (2014) conducted a CEA on five programs designed to decrease high school dropout. The five programs differed in a variety of ways but the researchers deemed all five evidence-based by the What Works Clearinghouse and used the same outcome metrics (number of participants who received a high school diploma or general equivalency degree) for evaluating effectiveness. Holland et al.'s (2014) results demonstrated the utility of CEA and how solely examining effectiveness can mislead practitioners. Two programs, Job Corps and National Guard Youth ChallenNGe (NGYG), demonstrated the largest effects. These programs graduated the greatest percentage of students in the treatment condition relative to the control condition. However, it was, in fact, the program Talent Search that proved to be, by far, the most cost-effective. The cost per graduated student for Job Corps (ICER = $131,140) and NGYG (ICER = $71,220) dwarfed that of Talent Search (ICER = $30,520); although Talent Search did not produce the same percentage of graduates as the other programs, the graduates it

did produce were relatively inexpensive. With this information, educators could decide to move other programs closer to the Talent Search model, cut programs that are extremely cost-inefficient, or examine what conditions cause some programs to outperform others. Considering the results in terms of cost per graduated student suggest that applying CEA in schools could uncover useful information for decision-makers regarding the programs and practices used.

Harbison and Hanushek (1992) conducted a study to examine the cost-effectiveness of eight possible interventions to increase literacy rates in Brazilian schools. These interventions included: (a) the provision of drinkable water, (b) the addition of desks to classrooms, (c) building new facilities such as bathrooms, (d) the provision of textbooks, (e) the provision of writing materials, (f) teacher training programs, (g) increasing teacher education, and (h) increasing teacher salaries. The researchers found that the single most cost-effective intervention to raise literacy rates among Brazilian students was the provision of textbooks (ICER = $0.26). This was followed closely by the provision of writing materials (ICER = $0.37), teacher training (ICER = $0.51), and the provision of drinkable water (ICER = $0.52). These are important results as the most effective intervention (without regard for CEA) was the building of new facilities and bathrooms. However, this intervention was so expensive (ICER = $1.79), that all four cost-effective interventions (i.e., textbooks, writing materials, teacher training, and clean water) could all be provided together for less.

In an early application of CEA, Quinn, Van Mondfrans, and Worthen (1984) conducted a CEA on a newly designed mathematics curriculum for Grade 6 students. As a measure of effect, the authors created a level of implementation measure, where classrooms fell on a spectrum between no implementation and total implementation of the new curriculum. The authors classified classrooms with low rates of implementation that primarily on the older curriculum as

business-as-usual (BAU). The authors then used multiple regression to calculate the expected gains on a mathematics criterion measure at each level of implementation. Quinn and colleagues also collected contemporaneous cost data at the time of the study. The authors found that the new mathematics curriculum—Goal-Based Education Management System Proficiency Mathematics (GEMS Math)—was less cost-effective than the BAU curriculum despite being more effective at increasing scores on the criterion measure. The ICER of not implementing GEMS math (i.e., administering the traditional math curriculum) was $194 while the ICER for GEMS math was $288. These results are of interest as they prove that newer curricula, interventions, or methods may not always be cost-effective, even if they demonstrate higher effect sizes.

Applications for Universal Screening

As Glover and Albers (2007) have noted, there are multiple considerations when selecting a universal screening measure. Perhaps the most important of these is diagnostic accuracy or classification accuracy. As such, much of the research on universal screening methods has been to quantify the SE, SP, PPV, NPV, AUC, and other metrics of accuracy for each screening method available. This has led to some confusion as often the relative differences in diagnostic accuracy between different methods are relatively small and difficult to interpret. In one example, Gersten and colleagues (2012) conducted a systematic review of screening methods for mathematics in middle school. They found 16 studies using 16 different screening methods and calculated the predictive validity of each. The authors found that no method had predictive validity < .34 and none had predictive validity > .79. Indeed, 6 of the 16 screening methods examined had predictive validity between .50 and .55, a small range. Without further information, it may be very difficult to select a single measure or even to rank-order all available measures based only on a very limited range of diagnostic accuracy indices.

Another important consideration identified by Glover and Albers (2007) is that of cost. If researchers cannot discriminate among screening methods based on diagnostic accuracy alone, the costs associated with conducting become valuable metrics to measure. While some researchers have attempted to address screening costs (Klingbeil et al., 2017; VanDerHeyden et al., 2017; VanDerHeyden et al., 2018), to date no systematic cost-analysis has been conducted. If researchers analyzed screening costs, significant variability could be found which could be leveraged to help select among screening methods. By combining both diagnostic accuracy and cost, CEA may be very beneficial in helping researchers and practitioners discriminate among screening methods.

Research Questions

1. Which individual screening measure is most cost-effective?
2. Which linear combination of multiple screening measures is most cost-effective?
3. To what extent does cost-effectiveness analysis reveal differences between screening methods relative to diagnostic accuracy?
4. To what extent are the findings from research questions 1 and 2 robust to variations in estimated costs?

Chapter Three

Methods

Extant Data

This study used extant data collected for a prior study and the methods section will detail how these data were collected and previously analyzed. Despite using extant data, this study uses novel analyses to answer entirely separate research questions. Data were originally collected as part of a study examining mathematics screening methods in middle school (Klingbeil et al.,

2019). The purpose of the study from Klingbeil and colleagues (2019) was to find which screening methods (CBM, MAP, the previous year's criterion exam) best predicted performance on a statewide criterion-referenced test (i.e., the Wisconsin Forward Exam). The district provided the researchers with data MAP and Wisconsin Forward Exam data. In addition to reporting the diagnostic accuracy indices (e.g., SE, SP) for the screening measures, the researchers also fit multiple regression models to create linear combinations of screening methods.

The diagnostic accuracy results from Klingbeil et al. (2019) were used in this dissertation to estimate the Diagnostic Odds Ratios associated with each screening measure and model. In addition, Klingbeil et al. (2019) did not address the cost of the screening measures at all, outside of reporting the time necessary to collect data using each approach. Using that information as a foundation, I collected additional information regarding the ingredients required to estimate the cost of each approach in an empirically sound way (Levin & McEwan, 2001).

Participants and Setting

The extant data used to inform this study were collected from students at a large suburban district in Wisconsin during the 2016-2017 school year. The Institutional Review Board at the University of Wisconsin – Milwaukee approved and managed the study. Participants in the original study included students in Grades 6, 7, and 8 at two suburban middle schools in Wisconsin. A total of 1,586 students participated in the study. The authors obtained passive consent from all participants. Students without consent to participate (n = 7) worked on other academic work while the CBM probes were administered. Approximately 69.9% of participants identified as White, 16.5% identified as Asian, 6.2% identified as Latinx, 4.3% identified as two or more races, 2.9% identified as Black, while less than 1% identified as either Native American, Alaskan Native, Native Hawaiian, or Other Pacific Islanders. Approximately 9.0% of students

qualified for free or reduced lunch. A small percentage (1.5%) of students were English language learners. Roughly 8.6% of students qualified for special education services.

Measures

This study examined three screening measures, the Wisconsin Forward Exam, the Measures of Academic Progress, and the AIMSweb Math Concepts and Applications CBM. A brief comparison of the three screening measures can be seen in Table 1.

Forward Exam. Klingbeil et al. (2019) used the Wisconsin Forward Exam as both the criterion measure and a possible screening measure in their study. Students complete the Wisconsin Forward Exam, a computerized test, in the spring of the academic year. The exam tests four content areas, but for this study, only students' Mathematics scores were used. The Forward Exam takes approximately 108 minutes to complete, per the exam publisher. The Wisconsin Forward Exam was designed to align with the Wisconsin Academic Standards in mathematics (Wisconsin Department of Instruction, 2016).

Information regarding the technical adequacy of the exam was published by the Wisconsin Department of Instruction (DPI). For the 2016 Wisconsin Forward Exam, the Wisconsin Department of Instruction assessed reliability by calculating Cronbach's alpha (α = .90 to .91), standard error of measurement for raw and scaled scores (2.72 to 2.82), classification consistency and accuracy, and inter-rater reliability for select items (Wisconsin DPI, 2016). Also, DPI provided evidence for construct and divergent validity. The Wisconsin Forward Exam in general—and the Mathematics content area in particular—was found to have adequate reliability and validity. The exam classifies students into four levels: (a) Below Basic, (b) Basic, (c) Proficient, and (d) Advanced. In the study from Klingbeil and colleagues (2019), the authors considered any student who was classified in the Below Basic or Basic levels as not proficient.

Ingredients. The Wisconsin Forward Exam assigns roles and responsibilities essential to test administration. Specifically, the Wisconsin Department of Public Instruction (DPI) divides these roles into two categories, district and school (Wisconsin DPI, 2020a). District roles include (a) the district assessment coordinator (DAC) and (b) the district technology coordinator (DTC). School roles include the (a) school assessment coordinator (SAC), (b) school technology coordinator (STC), and (c) test administrator/proctor. The DAC role is typically filled by an administration-level professional, usually with a specialty in academics, multi-tiered systems of support, or school psychology. The DTC is usually the district-level information technology (IT) director. At the school level, the SAC is normally a principal or vice-principal, the STC the school IT coordinator, and the proctor a classroom teacher.

The DAC coordinates all screening activities between the DTC and individual SACs (Wisconsin DPI, 2020a; Wisconsin DPI, 2020b; Wisconsin DPI 2020d). Responsibilities include scheduling training, disseminating materials, ensuring schools have resources such as laptops, and monitoring the administration of the Forward Exam. The SACs have very similar roles at the school level, with the additional responsibility of consulting with the STC (Wisconsin DPI, 2020e). The DTC is responsible for updating the Insight software to administer the Forward Exam, manage district-wide firewalls and IT protocols to ensure adequate access, and manages the individual STCs (Wisconsin DPI, 2000a; Wisconsin DPI 2020c). The STC performs necessary IT operations at the school level to ensure that hardware, software, and networks are ready for testing (Wisconsin DPI, 2020c). Finally, test proctors monitor students throughout the testing process, troubleshoot common issues, and report larger issues to SACs and STCs (Wisconsin DPI, 2020f). In addition to the personnel ingredients listed above, the Forward Exam also requires one laptop per student to be administered.

Measures of Academic Progress. The Measures of Academic Progress (MAP; Northwest Evaluation Association, 2011) was one of the screening measures evaluated in the study. The Northwest Evaluation Association (NWEA) designed the MAP as a computerized adaptive assessment intended to measure student achievement throughout the academic year. The MAP consists of multiple choice and short answer type items. The district administers the MAP in the fall, winter, and spring to track student progress. The approximate testing time is 45 minutes. The NWEA divided the MAP into two sections, Reading and Mathematics; only students' mathematics scores were used in the study. The Northwest Evaluation Association published extensive data suggesting that the MAP has adequate reliability and evidence of content, concurrent, predictive, and construct validity.

The MAP reports scores in Rasch Units (i.e., RIT scores), which are consistent across time allowing for the direct comparison of tests taken at different points during an academic year. MAP Mathematics scores range from 100 to 350. The Northwest Evaluation Association published a linking study that matches MAP scores with Wisconsin Forward Exam readiness levels (NWEA, 2017).

Ingredients. Although MAP testing is in a computer-adaptive format, there are significant personnel inputs that must be accounted for when conducting a CEA. The NWEA (2020a) provides guidance on assigning roles and responsibilities to schools and districts that is useful in conducting a CEA. Three roles are suggested to facilitate testing, while three more are recommended to conduct set-up and maintenance. Importantly, the NWEA allows for multiple individuals to fill each role, or a single individual to take on multiple roles. Despite this, the overall amount of personnel time required to conduct MAP testing should not change depending on exactly how schools assign roles; for instance, enrolling a fixed number of students into the

MAP platform should take the same amount of total time whether multiple people split the work or one person completes enrollment by themselves.

The NWEA advises schools and districts to assign a person or people to at least three roles to administer MAP testing: (a) school proctors (usually classroom teachers) who oversee students while they are testing or (b) a district assessment coordinator (DAC) who oversees the entire MAP testing process and enrolls students in the software, and (c) a school assessment coordinator (SAC) who monitors the testing status and ensures that all eligible students in a school complete MAP testing (NWEA, 2020a). In supporting roles, the NWEA suggests an additional two school personnel: (a) a system administrator who manages the MAP platform, including granting access to all other individuals, and (b) a data administrator who manages the enrollment roster (which students will complete MAP testing at a given time) at the district level.

Although the roles suggested by NWEA are generally straightforward, to conduct a CEA one should know which school personnel are likely to fill each role and have an estimate of the time commitment required. It is important to understand which personnel most probably will take on each role as there is a cost difference in personnel time depending on the individual's salary; it is less efficient to assign a school administrator to the role of a school proctor as their cost per hour is much greater than that of a general education teacher or even a paraprofessional. It is also important to estimate the time commitment required to fulfill each role as this estimate allows researchers to derive a total cost for a given personnel ingredient. The NWEA provides in-depth training guides which were used to estimate both which personnel are likely to fill each role and the time commitment required for each role (NWEA, 2020a; NWEA, 2020b; NWEA, 2020c). In addition to the personnel ingredients listed above, the MAP requires an annual license and one computer per student to be administered.

Math concepts and applications. Math problem-solving skills were also measured using AIMSweb Math Concepts and Applications (MCAP) probes (NCS Pearson Inc., 2012). MCAP probes were designed based on recommendations by the National Council of Teachers in Mathematics focal points and cover a wide array of content areas. MCAP problems require students to identify useful information (such as in a word problem) before performing the necessary computations to solve the problem. According to the published technical manual, MCAP probes have adequate alternate form reliability ($r \geq .86$ for all grades) and criterion validity ($r \geq .74$) between fall MCAP scores and the Illinois Standards Achievement Test (NCS Pearson Inc., 2012). In addition, MCAP scores were associated with adequate areas under the curve (AUC) values (> .88) when used to predict the Illinois Standards Achievement Test in Grades 6 through 8. AIMSweb provided cut-scores at both the 15th and 45th percentiles (NCS Pearson Inc., 2011. For MCAP probes, Klingbeil and colleagues (2019) selected cut scores based on the 45th percentile. These cut-scores were 11 for Grade 6, 10 for Grade 7, and 8 for Grade 8.

Ingredients. Of the screening measures evaluated in this study, MCAP appears to be the least resource-intensive due to its simplicity to administer. In terms of personnel, MCAP only requires one individual (typically a classroom teacher) to administer and score probes (Pearson Inc. 2012). An annual license is required to access individual probes. Paper is also a small but necessary material ingredient.

Table 1
Comparison of the Forward Exam, MAP, and MCAP CBM

	Forward Exam	MAP	MCAP
Administration Time	108 minutes	Approximately 45 minutes	8 minutes (Grade 6); 10 minutes (Grades 7 and 8)
Administration Format	Computer	Computer	Paper
Administration Frequency	Once per year	Thrice per year	As frequently as needed
Administration Period	Spring	Fall, Spring, Winter	When needed
Aligned Standards	Wisconsin Academic Standards	Common Core	Common Core
Use	Screening and criterion measure	Screening and progress monitoring	Screening and progress monitoring
Developer	Wisconsin Department of Public Instruction	Northwest Evaluation Association	Pearson Inc.

Procedure

The district regularly collected universal screening data using the Measures of Academic Progress (MAP). The MAP was administered district-wide three times per school year (fall, winter, and spring). The NWEA estimates that students in Grades 6 through 8 will take approximately 45 minutes to complete the mathematics section of the MAP (NWEA, 2018). Students in the district also completed an annual criterion measure of grade-level proficiency, the Wisconsin Forward Exam in the spring of 2017. DPI estimates the mathematics section of the Forward Exam will take approximately 105 minutes to complete for Grade 6 and Grade 7 students, and 115 minutes for Grade 8 students (DPI, 2019). Data from the 2016 and 2017 Forward Exams were collected. Klingbeil et al. (2019) used the 2017 Forward Exam data as the criterion measure and the 2016 Forward Exam data as one possible screening method. In other words, for a student in Grade 6, their results on the 2016 Wisconsin Forward Exam—completed while the student was in Grade 5—are used to predict current academic year criterion measure performance.

The district did not regularly conduct mathematics screening using AIMSweb curriculum-based measures (CBM), although these measures were used for diagnostic assessment and progress monitoring. As such, school personnel had experience administering CBM even though the context of universal academic screening was novel. The researchers and graduate students administered math concepts and applications (MCAP) and math computation (MCOMP) CBM to students in a class-wide setting. However, I only used data from the MCAP in this study, as AIMSweb recently discontinued the use of the MCOMP in the newest version of the measures (i.e., aimswebPLUS; NCS Pearson Inc., 2017). Math Concepts and Applications

probes take 8 minutes to complete for students in Grade 6 and 10 minutes to complete for students in Grades 7 and 8 (NCS Pearson Inc., 2017).

Graduate students and I conducted fidelity observations of every CBM administration. For MCAP, 98% of administration sessions were conducted with 100% fidelity (*M* = 99.84%). After CBM data were collected, the researchers and I hand-scored each probe. Of the 1,522 MCAP probes, approximately 21% were double scored to check for inter-rater reliability (IRR). IRR on the MCAP was 99.4%. The amount of time it took researchers to hand-score probes is included as a personnel ingredient in the CEAs for CBM. While the cost of scoring a single CBM probe is likely negligible, it must be noted that using CBM for universal academic screening requires scoring probes for an entire classroom, grade level, school, or district.

Prior Data Analysis

Klingbeil and colleagues (2019) used the data described above to obtain diagnostic accuracy indices for each screening measure using the vendor (or state) provided cut-scores, each screening measure using locally-derived cut-scores, and three linear combinations of screening methods. Klingbeil and colleagues (2019) calculated diagnostic accuracy for each screening measure or a linear combination of measures for each grade, which I used to calculate ICERs. Using the ingredients method and the CostOut toolkit, I estimated the total cost per student of each screening measure or combination of measures for each grade. Once diagnostic accuracy indices and screening costs have been determined, I calculated ICERs for each screening measure by grade.

All data analyses were separated by grade for two reasons. First, stratifying screening results by grade is consistent with applied screening practices in schools. Second, by separating

analyses by grade, any potential trends in DOR, costs, and/or ICERs that result as a function of grade can be identified.

We calculated descriptive and correlational statistics for each screening measure. We then evaluated the diagnostic accuracy of each measure (SE, SP, and DOR) using vendor-provided cut-scores to predict performance on the 2017 Wisconsin Forward exam. Third, we fitted two multiple regression models to predict performance on the 2017 Forward Exam. The first model included the 2016 Forward Exam and Fall MAP as predictors, while the second, full model added MCAP to the other two measures. Finally, we used the unstandardized predicted values from the models to conduct a Receiver Operator Characteristic (ROC) curve analysis.

A ROC curve is a graphical representation of the SE and SP of a screening measure at all possible screening thresholds. With any screening measure, researchers can either increase SE (while lowering SP) or increase SP (while lowering SE) by changing the at-risk/not-at-risk discrimination threshold on the screening measure (Hajian-Tilaki, 2013). In practical terms, researchers and screening measure developers want screeners to meet minimum standards of SE and SP ($\geq .90$ and $\geq .70$, respectively). If a screening measure has, for instance, a SE of .85, researchers can lower the discrimination threshold to increase the likelihood that students in general are categorized as at-risk (thereby ensuring that more students who are actually at-risk are captured by the measure). This action comes at a cost, however, as lowering the discrimination threshold (per our example above) also has the effect of decreasing SP (and vice versa if the discrimination threshold were increased).

By conducting a ROC analysis, researchers are able to determine what the discrimination threshold of a measure should be in order to achieve the desired threshold. Because the two linear combinations of measures in this study do not have vendor provided thresholds, ROC

curve analyses are necessary to select a threshold where SE ≥ .90 and SP ≥ .70. Once the discrimination threshold of a linear combination of measures is selected, indices of diagnostic accuracy (such as TP, TN, FP, FN, and DOR) can be derived.

Missing data. In the extant data set, approximately 6.87% of students were missing data on the 2016 Forward Exam, 4.04% on MCAP, 2.08% on the 2017 Forward Exam, and 0.25% on the fall MAP. Results of Little's (1988) missing completely at random (MCAR) test did not support the assumption that data were MCAR. I, therefore, assumed that the data were missing at random (MAR)—this assumption is not empirically testable but follows suggested procedures for multiple imputation (Peugh & Enders, 2004). Indeed, as Graham (2009) has noted, even if the assumption of MAR is violated, multiple imputation is still preferred to list-wise deletion. Although a violation of MAR would affect the multiple imputation procedure, it would also affect list-wise deletion, likely to a greater extent. Klingbeil et al. (2019) imputed 20 data sets using student gender, ELL status, SES, special education status, and race/ethnicity as predictor variables. Math achievement data were used both as predictor and imputed variables. The 20 data sets were pooled with the resulting imputed set used for all analyses. Multiple imputation has been shown to introduce less bias into the analyses than list-wise deletion (Peugh & Enders, 2004).

Analytic Plan for Research Question 1

Diagnostic Odds Ratios. In intervention research, researchers conducting CEA use effect sizes such as standardized mean differences, which do not apply to universal screening. As there is no widely accepted metric for effect size in universal screening, I used diagnostic odds ratio (DOR) values as an effect size for the CEA. In universal screening research, DOR values are often used as holistic metrics of overall screener performance (Glas et al., 2003). DOR values

range from zero to infinity and represent the likelihood that a screening measure correctly classifies a student as at-risk or not-at-risk. The formula for calculating DORs is:

$$DOR = \frac{TP/FN}{FP/TN} \tag{1}$$

Additionally, a 95% confidence interval for DORs can be calculated by obtaining the standard error (SE) of the log of the DOR, obtaining a confidence interval using the SE, then calculating the antilog of the resulting expression (Glas et al., 2003):

$$SE(\log DOR) = \sqrt{\frac{1}{TP} + \frac{1}{TN} + \frac{1}{FP} + \frac{1}{FN}} \tag{2}$$

$$\log DOR \pm 1.96SE(\log DOR) \tag{3}$$

A screening measure with a DOR value of exactly 1.0 performs no better than random chance—it has no diagnostic value over and above randomly selecting students for intervention. As the diagnostic accuracy of a screening measure increases or strengthens, DOR values increase (i.e., larger DOR values indicate better screening accuracy). A screening measure with a DOR value below one represents a unique scenario; the screening measure so reliably incorrectly categorizes students as at-risk or not-at-risk that diagnostic accuracy can be improved simply by inverting the measure. DORs are closely related to Youden's index, another summary metric of

diagnostic accuracy, but are more easily interpreted as they do not require linear translation (Glas et al., 2003).

DORs have been used for some time in the literature on universal screening in schools and are widely accepted as valid and useful measures of screening accuracy (Kilgus, Methe, Maggin, & Tomasula, 2014). There are, however, some criticisms of DORs found in the medical literature, particularly that the statistic is not a good predictor of the effectiveness of a screening measure or test for a single individual (Pepe, Janes, Longton, Leisenring, & Newcomb, 2004). In short, medical researchers argue that even tests with very large DORs tend to not perform particularly well when used to classify a single patient due to the variety of covariates that can be present; a screening measure for heart disease with a large DOR may still not perform well when used on a person who has a genetic predisposition to heart disease, for example. These criticisms, while valid, may not apply to research in education or psychology, however. While a universal screening measure based on logistic regression that incorporates a wide variety of variables would almost certainly perform extremely well in education, this approach is not generally used as part of universal screening conducted in schools.

There are essential elements of DORs that make them an appropriate choice for an effective size when evaluating the diagnostic accuracy of screening tools and procedures. First, DOR values are interval data represented on a spectrum with an infinite number of possible values. Second, DOR values increase as a measure performs better, similar to how measures of effect size increase as interventions perform better. Finally, DORs are already widely used to compare different screening measures as effect sizes are used to compare different interventions. There are, however, some key differences between DORs and traditional effect sizes that should be noted. First, DOR values are not true ratio data as a DOR of zero is, practically, meaningless.

A screening measure with a DOR < 1.0 will incorrectly classify students at a rate worse than random chance. If taken to its logical conclusion, a screening measure with a DOR = 0.0 would incorrectly categorize students 100% of the time. Most schools treat screening decisions as a dichotomous classification, the decision to intervene or to withhold intervention, a school psychologist would only have to perform the inverse action recommended by this hypothetical screening measure to correctly provide intervention for every student-at-risk. This thought experiment illustrates the problem with rationalizing DOR values below 1.0. With this consideration in mind, it is difficult to imagine a screening measure developed in an applied setting with a DOR < 1.0. Even a mathematics screening measure that consisted solely of geography questions would very likely have a DOR > 1.0 as students who perform better in geography are also likely to perform better in mathematics. Despite some minor limitations, there appears to be no reason why DORs could not be used in a CEA.

Ingredients. After calculating diagnostic accuracy, I used the CostOut Toolkit and the ingredients method (Levin & McEwan, 2001) to determine costs. Initially, published standards, procedures, and supplemental materials for the screening tools I evaluated were used to create a preliminary list of ingredients. The ingredients for each screening measure were considered without respect for ingredients already defined for other measures. For instance, a laptop of some type is required to take both the Measures of Academic Progress and the Wisconsin Forward Exam; the cost of the laptop was factored in for each screening measure, even though in practice only one laptop could be used for multiple screening measures. This decision was made so as not to bias the results of the CEA by arbitrarily selecting which screening measures should bear the costs of non-consumable items such as laptop computers or license fees. In other words, by

factoring in ingredients independently for each screening measure, one can control for these costs to make a direct comparison between measures possible.

Under the facilities category, I presumed no cost for any of the screening measures. Although there are certainly some costs that are associated with classroom use (e.g. heating, electricity, cleaning), researchers suggest that facility costs be conceptualized in terms of creating new space or renting/buying existing space (Levin & McEwan, 2001) As existing classrooms are usually assigned to a particular teacher or academic subject, there is likely no additional cost for using a room for academic screening instead of, for instance, general math instruction. The cost of janitorial staff, heating/air conditioning, or electricity—for example—would not be expected to vary as a function of what academic activity is being conducted in a classroom. For this reason, facilities costs are not included in the analyses.

In terms of hardware such as laptops, costs were estimated by calculating percent usage, not the cost of purchase. As mandatory state testing is largely computer-based, it can be assumed that districts already have hardware available. Despite this fact, it would be incorrect to assume that there is no cost associated with using hardware that has been previously purchased. As noted by Levin and McEwan (2001) there are real costs with using non-consumable materials such as computers. There are maintenance costs (computers degrade over time and usage) as well as opportunity costs (a computer used by one student cannot be used by another simultaneously) that need to be considered. The CostOut Toolkit allows the user to account for this by defining several ingredient conditions. First, I estimated the total cost of the computer or laptop based on the CostOut Toolkit's database of ingredients. Then the expected number of years that a laptop or computer can be expected to be used before being replaced was defined. Finally, I calculated the percentage of time that the laptop or computer will be used out of an entire school year. By

calculating these values, one can estimate how much of the cost of a laptop or computer can be attributed solely to its usage for one universal screening measure. An example with the actual costs of buying and maintaining a Chromebook for use in MAP screening can be seen below: with the given inputs: (a) the cost of a new Chromebook is $166.00, (b) the annual license fee is $25.00, (c) the annual damage protection fee is $17.25, (d) the cost of a carrying case is $19.00, and (e) the lifetime of the laptop is four years. Note that these costs were provided by the director of information technology at a district like that which participated in the study.

$$Usage_{screening} = \frac{Hours_{screening}}{Hours_{school\ year}} = \frac{.75}{1{,}260} = 0.0006 \quad (4)$$

$$Cost_{Computer} = Cost_{Chromebook} + Cost_{Carrying\ Case} + (Cost_{License} \times Years_{Chromebook})$$

$$+ \left(Cost_{Damage\ Protection} \times Years_{Chromebook}\right) = 166.00 + 19.00 + (25.00 \times 4) + (17.25 \times 4) = 354.00$$

$$Cost_{ingredient} = \left(\frac{Cost_{computer}}{Years_{computer}}\right) \times Usage_{screening} = \left(\frac{354.00}{4}\right) \times 0.0006 = 0.0531$$

As shown above, the actual cost for the use of a laptop for one-half of a school day equates to only about $0.25 per student. This example is useful as it can be used to demonstrate the applied value of CEA. Without conducting a CEA, some may be tempted to consider a computer an expensive ingredient in screening since computers often cost hundreds of dollars. This assumption is largely incorrect as hardware is not a consumable ingredient, meaning that it can be used more than once and for purposes other than universal screening. If hardware and other non-consumable ingredients are incorrectly classified as consumable, it would mean that schools would have to buy laptops solely for universal screening and then dispose of them after a single use. This is, of course, inconsistent with how laptops and computers are used in schools. Other non-consumable ingredients include licenses for statistical software, facility usage, etc.

Not all costs could be estimated using the CostOut Toolkit. Examples of ingredients without costs defined by CostOut include license fees for AIMSweb or the MAP, training costs such as time spent in instructional seminars, etc. Whenever possible, I obtained these ingredient costs directly from the publisher of a particular screening measure. When costs could not be estimated either through the CostOut Toolkit or from the measure publishers, I calculated costs based on average market rates and interviews with school personnel in charge of purchasing. As an example, Chromebooks require a yearly license fee to operate as shown above. The cost of this license cannot be found in either the CostOut toolkit or from the manufactures of Chromebooks. In this case, I contacted the director of information technology at a large suburban school district in Ohio who provided me the exact prices the district pays to supply its students with Chromebooks. These costs are similar to those found in the district which participated in this study. When prices cannot be found in the CostOut Toolkit, obtained from a publisher, or estimated based on expert knowledge, best practice is to look up actual advertised prices of

products sold in popular marketplaces (Levin & McEwan, 2001). Indeed, this is one way that researchers built the CostOut database—for instance, the price of paper in CostOut is based on the price of a ream from Staples in 2018.

Another cost that can be defined using the CostOut toolkit is classroom teacher or school psychologist time. However, several conditions needed to be defined before a cost could be estimated. First, I calculated the total time (in minutes) required by school personnel to administer a universal screening measure. The total time may include the time needed for training, administering measures, scoring measures, etc. Whenever possible I used exact times provided by test publishers—for instance, the NWEA estimates that the MAP will take approximately 45 minutes for students to complete. When the time was not provided (as in the case of time to score a CBM), I made estimates based on practical considerations and experience. For example, I estimated that it takes a teacher or school psychologist approximately one minute to score a CBM probe. This number is derived from the actual amount of time it took researchers to score CBM probes in previous research (Klingbeil et al., 2019).

Once the amount of time was estimated, other variables were defined in the CostOut Toolkit. The toolkit allows the user to estimate the cost of school personnel time using several factors including the number of years of experience, geographic location, and even the cost of employee benefits such as health insurance. For instance, a high school teacher with 15 years of experience at the national average compensation rate costs $42.59 per hour, whereas a high school teacher with only 5 years of experience at the national average compensation rate costs only $30.33. It is this ability to disaggregate ingredient costs that make the CostOut Toolkit so powerful, but also slightly cumbersome. To ensure that costs are standardized across screening measures, I selected ingredients that fulfill the following conditions: (a) national average, (b) 15

years of experience, and (c) no special certifications or qualifications. The average benefits package for teachers and school psychologists was also calculated using the CostOut Toolkit and was combined with estimated salaries to increase accuracy. National averages were used instead of regional averages to increase the external validity of the study. I selected 15 years of experience as the national average teacher experience (in years) is 13.8 in public schools (National Center for Education Statistics [NCES], 2012).

Once all diagnostic accuracy indices and ingredient costs were defined, the final inputs to consider were class and district size. In this study, it is important to understand when to apply screening costs to individual students, classrooms, schools, or districts. For example, the Wisconsin Forward Exam requires both an STC and DTC for administration. When an STC performs necessary network diagnostics, only students in that school benefit, and therefore the costs of an STC need to be spread over only those students. Alternatively, when a DTC updates the Insight software, every student in the district benefits, and the cost of a DTC's time needs to be spread over every student enrolled in the district. As such, the number of students at the class, school, and district levels must be defined. District size was obtained through public records; the district that participated in the study serves approximately 7,600 students. As class size in the district was not readily available, it was estimated using NCES data. The NCES regularly publishes the average classroom size throughout the U.S. In Grades 6 through 8, the average class size for a general education classroom is 25.5 (NCES, 2012b). I rounded this up to the nearest whole student, for an average class size of 26.

After diagnostic accuracy indices, ingredient costs, and class sizes were defined, the process of calculating ICERs is relatively simple. Ingredient costs were added, with the resulting sum divided by the number of students be screened, and the resulting quotient then divided by

the measure's DOR value. The resulting ICERs were compared to understand which universal screening measures are the most cost-effective for each grade in the data set.

$$\frac{\left(\frac{Ingredient\ costs}{Number\ of\ students\ screened}\right)}{DOR} \quad (5)$$

Analytic Plan for Research Question 2

In order to evaluate the diagnostic accuracy of multiple screening measures, two and three individual measures were combined using multiple linear regression similar to procedures used by Catts et al. (2001) and Nelson et al. (2016). First, two linear regression models were built, both predicting scores on the 2017 Forward Exam. The first model included the 2016 Forward Exam and the Fall MAP as predictors. The second model included the 2016 Forward Exam, Fall MAP, and MCAP as predictors. Once the models were built, the resulting unstandardized predicted values (of 2017 Forward Exam scores) were saved for all participants.

The unstandardized predicted values were then used to create a ROC curve. A ROC analysis requires both a dichotomous criterion variable and a continuous predictor variable. The unstandardized predicted values discussed above were saved and used as the continuous predictor variable. The dichotomous criterion variable was created by categorizing students as at-risk or not-at-risk based on their predicted values. The ROC curve analysis first selects an arbitrary discrimination threshold, then applies this threshold to participants predicted scores on the 2017 Forward Exam, resulting in two groups of participants categorized as at-risk or not-at-risk. The predicted classifications are compared to students' actual classification on the 2017 Forward Exam (i.e., criterion measure) and SE and SP are calculated (using the TP, TN, FP, and FN data derived from comparing a participants predicted condition with their actual condition).

This process is repeated for every possible value of the discrimination threshold, resulting in a graphical plot. These ROC curve plots were used to select discrimination thresholds where SE ≥ .90. The resulting TP, TN, FP, and FN produced by selecting this particular discrimination threshold were used to calculate DOR for each model respectively.

In terms of costs, a two-step process was used. First, the total screening costs (i.e., the total costs of all individual ingredients) for each measure in the linear regression was added together. Second the cost of two additional ingredients, school psychologist time and an SPSS license fee, required to conduct the multiple linear regression and ROC curve analyses were calculated. The amount of time to conduct all analyses was estimated based on the amount of time it took me to complete them. This amount of time was multiplied by the per minute ingredient cost of a school psychologist to obtain a total ingredient amount. Finally, the cost of an annual SPSS license was divided by the number of participants in the study to obtain a total ingredient cost. These two ingredient costs were added to the sum of the costs of the screening measures themselves to obtain a final cost per student when using combinations multiple screening measures. More detailed ingredient costs for Research Question 2 can be found in Tables A.6 through A.10 in the Appendix.

Analytic Plan for Research Question 3

One problem schools and districts encounter when selecting a screening approach from several screening measures is a general heterogeneity in diagnostic accuracy estimates. As most researchers agree on minimum standards of sensitivity and specificity, a relatively high floor is set for screening measures to be considered valid for high-stakes decision making. Because of this high floor, screening methods tend to cluster together with relatively high levels of diagnostic accuracy, making it difficult to determine which is the most effective for a school or

district. An example of this effect can be seen in data collected by Klingbeil et al. (2019). When using locally derived cut scores to ensure that every screening method had minimally acceptable sensitivity (i.e., SE ≥ .90), Klingbeil and colleagues found little variation in specificity.

To answer research question four, I calculated the coefficient of variation (C_V) for the three individual screening measures in the study in terms of both DOR and ICER. The C_V is a measure of variability that controls for the unit of measurement among a set of numbers (Abdi, 2010). It can be defined as the ratio of the standard deviation to the mean.

$$C_V = \frac{SD}{M} \tag{6}$$

There were three screening approaches evaluated in each grade (9 total). Only individual screening measures were evaluated. The two linear combinations of screening measures were not evaluated due to issues with intercorrelation; because the two linear models both included the same costs for the Forward Exam and the MAP, the standard deviations of the ICERs produced these two models would be more similar compared to individual measures with their own unique costs. To compute the C_V, I first calculated the standard deviation of DOR and ICER values across all grades and screening methods. I calculated the mean DOR and mean ICER across all grades and individual screening methods. I then estimated the standard deviation of DORs and ICERs (respectively) by grade. These value were then divided by the mean DOR and mean ICER to obtain a C_V.

For this study, the value of C_V is its ability to be used to compare the variance of DOR and ICER, two metrics with different units of measure. If the C_V is larger for ICER than DOR, we can conclude that the results of the CEA provide additional information to differentiate between screening measures.

Analytic Plan for Research Question 4

The ideal method in conducting a CEA is to collect detailed cost information during the actual implementation of an intervention, program, or curriculum. However, a post hoc CEA can be performed with careful consideration towards accurately estimating costs. In this study, cost estimates can be categorized into three tranches: (a) low-level assumptions, (b) mid-level assumptions, and (c) high-level assumptions.

Low-level assumptions include costs that can be estimated with a high degree of accuracy and confidence. Costs for ingredients such as the paper required for MCAP, the cost of an annual SPSS license needed to conduct multiple linear regression, or teacher time to complete the MAP training module involve low-level assumptions; the cost of paper was obtained directly from a wholesaler, the cost of an SPSS license was drawn from IBM's product pricing, and the MAP training module is computerized and takes exactly 60 minutes.

Mid-level cost assumptions are less precisely defined than low-level assumptions, but still have a moderate to high degree of accuracy. An example of a mid-level cost assumption would be the price of a laptop to administer the Wisconsin Forward Exam or the MAP. One of the assumptions required to determine the cost of a laptop is the lifetime of the technology; a laptop that lasts three years is inherently more costly to use per minute than a laptop that lasts four years. For this example, I spoke with the Director of Technology at a large suburban school district in Ohio. The Director of Technology stated that, at their district, laptops are replaced every four years (laptops are bought for a cohort of students in Grade 1 and replaced in Grades 5 and 9). As such, in this study, it was assumed that the life of a laptop was four years (see the exact equation for laptop cost on p. 37). While we can say with a moderate to high degree of confidence that assumed laptop cost is very similar to real laptop cost, it is possible that different

districts have slightly different guidelines on which laptops to purchase, how long they are used before being replaced, etc. As such, laptop cost is considered a mid-level assumption.

High-level cost assumptions are most likely to affect the validity of the analysis results. These costs are estimated from sources such as published documents (for instance, technical manuals), direct interviews with district staff (for example, the Multi-Tiered Systems of Support District Leader at the district where data were gathered), or personal experience (for instance, the amount of time required to score a single MCAP probe was estimated based on the writer's experience scoring probes during data collection). Despite this, costs—especially those related to human activities such as data entry—can vary greatly as a function of effort, technology, experience, etc. The costs in this study most susceptible to these factors are the personnel ingredients at the school and district level including inputs such as DTC preparation time for the Wisconsin Forward Exam or SAC time to organize and manage the administration of the MAP across an entire school. Put simply, there is some variability in the efficiency of school and district staff to prepare for and administer universal academic screening. To account for these high-level cost assumptions, two sensitivity analyses were conducted. In the first, personnel time inputs that are not explicit (such as proctor time, which is static as a function of published testing time) were increased by 50%. In the second, those personnel time inputs were decreased by half.

CHAPTER FOUR

Results

In this chapter, I report results for research questions 1, 2, 3, and 4. Research question 1 asked which individual screening measure is the most cost-effective as determined by a CEA. Research question 2 asked which linear combination of screening methods is most cost-effective. Research question 3 concerns the relative variability between screening measures before and after conducting a CEA to determine if CEA provides useful discriminatory information for selecting measures. Finally, research question 4 seeks to determine whether the results for research questions 1 and 2 are robust to changes in estimated costs. Diagnostic odds ratios (with 95% confidence intervals), screening measure cost per student, and ICERs (with 95% confidence intervals) stratified by are reported in Table 2. The adjusted ICERs derived to answer research question 4 can be found in Table 3. Specific ingredient costs and quantities for each screening model and subsequent CEA can be found in the Appendix.

Research Question 1

Overall, the most cost-effective screening measure was the Wisconsin Forward Exam from the previous year (ICER *M* = $0.238; range = $0.141 to $0.402). The next most cost-effective screening measure is the linear combination of students' scores on the Wisconsin Forward Exam and Fall MAP screeners (average ICER = $0.583; range = $0.329, $1.035). The ICER for the MAP alone was highly similar (average ICER = $0.612; 95% CI = $0.368, $1.019) to the linear combination of the Wisconsin Forward Exam and the MAP.

Curriculum-based measurement resulted in much higher ICERs. The MCAP alone produced an average ICER of $0.800 (range = $0.436, $1.47), making it over three times less

cost-effective than the Wisconsin Forward Exam as a screening measure. When added to the linear combination of Wisconsin Forward Exam score and MAP scores, MCAP decreases the overall diagnostic accuracy of the model while simultaneously increasing cost, leading to the highest average ICER of $0.848 (range = $0.481, $1.493). Full results broken down by screening measure and stratified by grade can be seen below.

Table 2

DOR, Cost per Student, and ICER by Screening Method and Grade

Method	Grade	DOR [95% CI]	Cost/Student	ICER [range]
Prior Year Forward Exam				
	6	20.938 [12.654, 34.670]	$7.57	$0.361 [$0.218, $0.598]
	7	41.580 [24.333, 71.052]	$7.57	$0.182 [$0.106, $0.311]
	8	47.047 [27.133, 81.578]	$8.01	$0.170 [$0.098, $0.295]
MAP				
	6	31.936 [18.766, 54.379]	$19.40	$0.608 [$0.357, $1.034]
	7	26.244 [16.204, 42.505]	$19.40	$0.739 [$0.456, $1.197]
	8	39.582 [23.483, 66.716]	$19.40	$0.490 [$0.291, $1.827]
MCAP				
	6	15.878 [8.038, 31.364]	$10.99	$0.692 [$0.350, $1.367]
	7	9.835 [5.733, 16.872]	$11.08	$1.126 [$0.657, $1.932]
	8	19.051 [9.893, 36.687]	$11.08	$0.582 [$0.302, $1.120]
Forward + MAP				
	6	40.267 [22.264, 72.828]	$27.82	$0.691 [$0.382, $1.250]
	7	43.759 [24.946, 76.759]	$27.82	$0.636 [$0.362, $1.115]
	8	66.772 [38.172, 116.908]	$28.27	$0.423 [$0.242, $0.751]
Forward + MAP + MCAP				
	6	39.018 [21.893, 69.708]	$38.81	$0.995 [$0.557, $1.777]
	7	42.751 [24.384, 74.951]	$38.90	$0.910 [$0.519, $1.595]
	8	61.618 [35.513, 106.913]	$39.35	$0.639 [$0.368, $1.108]

Note. The range for ICERs are derived by multiplying the cost per student value by the upper and lower 95% CI values for each DOR.

Prior Year Forward Exam. Relative to other individual screening measures, the prior year Forward Exam demonstrated the greatest diagnostic accuracy—as measured by diagnostic odds ratios—for all students in Grades 7 (DOR = 41.580; 95% CI = 24.333, 71.052) and 8 (DOR = 47.047; 95% CI = 27.133, 81.578). The prior year Forward Exam did not perform as well for Grade 6 students (95% CI = 20.938; range = 12.645, 34.670) relative to the MAP, which demonstrated the greatest diagnostic accuracy of any individual measure in grade six. Overall, the prior year Forward Exam had an average DOR of 36.523 (95% CI = 21.370, 62.433) across grades.

The Wisconsin Forward Exam had lower costs per student than any individual measure or combination of measures in every grade. For Grades 6 and 7, the cost per student was $7.57, whereas in Grade 8 the cost per student was $8.01. The increase in costs in Grade 8, relative to other grades, is due to the increased administration time of the measure for Grade 8 students. The Wisconsin Forward Exam administration time is ten minutes longer for Grade 8, leading to increased costs in both materials (laptop usage) and personnel (teacher proctoring time). It should be noted that, while both materials and personnel costs increased with the additional administration time, the vast majority of the greater cost was due to teacher proctoring time (the additional laptop usage only amounted to about a penny per student increase). The average cost per student across all grades to administer the Wisconsin Forward Exam was $7.76.

Due to the relatively high diagnostic accuracy of the prior year Forward Exam, combined with its low cost, the ICERs for each grade were the lowest of any individual or combination of screening measures. In Grade 6 the ICER was $0.361 (range = $0.218, $0.598), in 7 $0.182 (range = $0.106, $0.311), and in 8 $0.170 (range = $0.098, $0.295). Overall, the average ICER for the Wisconsin Forward Exam across grades was $0.238 (range = $0.141, $0.402). These

results suggest that the most cost-effective method of conducting universal academic screening in middle school would be using data from the previous year's criterion measure.

MAP. In terms of single screening measures, the MAP had the greatest diagnostic accuracy in Grade 6 (DOR = 31.936; 95% CI = 18.776, 54.379). It underperformed relative to the Wisconsin Forward Exam in Grades 7 (DOR = 26.224; 95% CI = 16.204, 42.505) and 8 (DOR = 39.582; 95% CI = 23.483, 66.716), but had greater diagnostic accuracy in all grades than MCAP CBM probes. For all middle school grades, the MAP had an average DOR of 32.587 (95% CI = 19.484, 54.533)

The MAP was the most expensive screening method tested in this study, with an average cost per student across grades of $19.40, well over twice the cost of the Wisconsin Forward Exam and nearly twice the cost of MCAP. The high cost of the MAP is almost entirely due to the individual licensing fee required to administer it to each student; the annual MAP license accounted for approximately 70% of the total cost per student. The MAP was the only screening measure evaluated that did not have an average cost per student that varied as a function of grade, making average costs more meaningful.

Although the MAP performs well as an individual screening measure (i.e., meeting established standards of diagnostic accuracy), its expense makes it a less cost-effective screening method than the previous year's Wisconsin Forward Exam. In Grade 6 the ICER was $0.608 (range = $0.357, $1.034), in 7 $0.739 (range = $0.456, $1.197), and in 8 $0.490 (range = $0.291, $0.826). Overall, the average ICER for the MAP across grades was $0.612 (range = $0.368, $1.019). In terms of cost-effectiveness, using MAP scores alone for universal screening represents an over two-fold increase in cost per unit of diagnostic accuracy (i.e., DOR) as opposed to the Wisconsin Forward Exam.

MCAP. The MCAP CBMs performed markedly worse than all other screening methods evaluated on screening accuracy. The MCAP did not meet the minimum suggested requirements for universal academic screening (i.e., sensitivity ≥ 0.90 and specificity ≥ 0.70) in any grade. Regarding DOR, the MCAP underperformed all other measures and methods in Grades 6 (DOR = 15.878; 95% CI = 8.038, 31.364), 7 (DOR = 9.835; 95% CI = 5.733, 16.872), and 8 (DOR = 19.051; 95% CI = 9.893, 36.687). Across grades, the average DOR for MCAP was 14.921 (95% CI = 7.888, 28.308).

Although MCAP did not demonstrate adequate diagnostic accuracy, it was relatively inexpensive to administer, costing only about half of the MAP. The average cost per student across grades was $11.05. For Grade 6, the cost per student was $10.99, while in Grades 7 and 8 the cost per student was $11.08. As with the Wisconsin Forward Exam, the difference in cost between grades was a function of increased administration time. Much like the MAP, the MCAP licensing fee accounted for a large amount of the cost per student. While the average cost per student across grades was $11.05, the license fee was $6.50, or 59% of the total.

The average ICER for MCAP in Grades 6, 7, and 8 was $0.800 (range = $0.436, $1.473). The MCAP was less cost effective than any other measure or combination of measures in grades 6 (ICER = $0.692; range = $0.350, $1.367), 7 (ICER = $1.126; range = $0.657, $1.932) and 8 (ICER = $0.581; range = $0.302, $1.120). Despite the relatively low cost of administering MCAP and other CBM, the low diagnostic accuracy of these probes made them the least efficient method of screening analyzed in this study.

Research Question 2

Using two multiple linear regression models predicting the criterion measure (i.e., the 2017 Wisconsin Forward Exam), I combined information from the Forward Exam and MAP and then the Forward Exam, MAP, and MCAP into a single predicted 2017 Forward Exam score. Full regression results for each model are shown in Table 3. I used the predicted values as a composite screening score to determine each student's risk status. Risk status was derived using unstandardized predicted values in a Receiver Operator Curve analysis, allowing for the selection of a cut score where sensitivity ≥ .90. I then created a dichotomous risk variable based on this threshold and used that to calculate DOR for each screening model.

Forward + MAP. The first model combined the prior year Forward Exam and the MAP. This linear combination resulted in greater diagnostic accuracy than any individual screening measure in isolation. When combining the prior year Forward Exam and the fall MAP, the DOR across grades was 50.266 (95% CI = 28.461, 88.832). The DOR for Grades 6, 7, and 8 were 40.267 (95% CI = 22.264, 72.828), 43.759 (95% CI = 24.946, 76.759) and 66.772 (95% CI = 38.172, 116.908), respectively.

In terms of costs associated with administering multiple screening measures, it is important to account not only for the combination of costs for each screening measure but the time and materials required to build the multiple linear regression models themselves. The cost per student of the linear combination of screening methods then includes the cost of each measure alone, plus the cost of school psychologist time to run the analyses, as well as the cost of a commonly used statistical analysis program (SPSS 27.0). The average cost per student across grades for Wisconsin Forward + MAP was $27.97. In Grades 6 and 7, the cost per student was $27.82, while in Grade 8 it was $28.27.

Table 3

Multiple linear regression results by grade and model

Grade	Model	Predictor	*B*	95% CI	*SE*	*t*	*p*
6	Forward + MAP	Intercept	-67.664	[-105.176, -30.152]	19.132	3.537	<.001
		2016 Forward Exam	0.465	[0.355, 0.575]	0.056	8.318	<.001
		Fall MAP	1.809	[1.486, 2.133]	0.165	10.978	<.001
	Forward + MAP +MCAP	Intercept	-10.789	[-62.915, 41.337]	26.578	0.406	.685
		2016 Forward Exam	1.565	[1.211, 1.918]	0.180	8.681	<.001
		Fall MAP	0.444	[0.334, 0.554]	0.056	7.930	<.001
		MCAP	0.594	[0.227, 0.962]	0.188	3.169	.002
7	Forward + MAP	Intercept	-69.435	[-108.308, -30.561]	19.832	3.501	<.001
		2016 Forward Exam	0.468	[0.353, 0.584]	0.059	7.959	<.001
		Fall MAP	1.799	[1.464, 2.134]	0.171	10.522	<.001
	Forward + MAP +MCAP	Intercept	-20.072	[-73.500, 33.356]	27.257	0.736	.461
		2016 Forward Exam	1.637	[1.285, 1.990]	0.180	9.109	<.001
		Fall MAP	0.436	[0.319, 0.554]	0.060	7.276	<.001
		MCAP	0.560	[0.145, 0.975]	0.212	2.646	.008
8	Forward + MAP	Intercept	-63.792	[-101.823, -25.762]	19.403	3.288	.001
		2016 Forward Exam	0.359	[0.268, 0.451]	0.046	7.736	<.001
		Fall MAP	2.048	[1.737, 2.359]	0.159	12.919	<.001
	Forward + MAP +MCAP	Intercept	-14.589	[-69.155, 39.977]	27.838	0.524	.600
		2016 Forward Exam	1.852	[1.506, 2.198]	0.177	10.487	<.001
		Fall MAP	0.345	[0.253, 0.436]	0.047	7.387	<.001
		MCAP	0.561	[0.117, 1.005]	0.226	2.479	.013

Note. SPSS multiple linear regression utilizing imputed data does not produce β values

Despite the increased diagnostic accuracy obtained by combining the Wisconsin Forward + MAP, the costs of administering both measures and building the multiple linear regression models made this method much less cost-effective than simply administering the Wisconsin Forward Exam alone. The ICER across grades for Wisconsin Forward + MAP was $0.583 (range = $0.329, $1.035). In Grade 6 the ICER was $0.691 (range = $0.382, $1.250), in Grade 7 the ICER was $0.636 (range = $0.362, $1.115), and in Grade 8 the ICER was $0.423 (range = $0.242, $0.741).

Forward + MAP + MCAP. The linear combination of the Wisconsin Forward Exam, MAP, and MCAP resulted in greater average diagnostic accuracy (DOR = 47.795; 95% CI = 27.245, 83.857) than any individual screening measure, but performed worse than the linear combination of only the Wisconsin Forward Exam and MAP (DOR = 50.266). The inclusion of MCAP resulted in less diagnostic accuracy as, in this study, MCAP was found to be a substandard screening method. This result suggests that the quality of the screening method may be more important when making screening decisions than the number of screeners administered. The DOR for the linear combination of Forward + MAP + MCAP varied substantially as a function of grade; DOR for the linear model was 39.018 (95% CI = 21.893, 69.708) for Grade 6, 42.751 (95% CI = 24.284, 74.951) for Grade 7, and 61.618 (95% CI = 35.513, 106.913) for Grade 8.

As three individual screening methods are administered, this linear combination was the costliest approach evaluated in this study. After factoring in school psychologist time to conduct the analyses and the cost of the statistical software use, this linear combination of screening methods resulted in an average cost per student across all grades of $39.02. In Grade 6 the cost per student was $38.81, in Grade 7 $38.90, and in Grade 8 $39.35. These costs represent a five-

fold increase compared to the Wisconsin Forward Exam alone and are almost twice the cost of the MAP alone. Relative to the linear combination of Forward + MAP, the full model has an average cost increase of 40% and an average DOR decrease of 5%.

Indeed, the linear combination of Forward + MAP + MCAP had the highest average ICER of any screening method in this student. Across Grades 6, 7, and 8, this linear combination had an ICER of $0.848 (range = $0.481, $1.493). In Grade 6, the ICER was $0.995 (range = $0.557, $1.777), in Grade 7 $0.910 (range = $0.519, $1.595), and in Grade 8 $0.639 (range = $0.368, $1.108). The linear combination of all three individual screening measures was the least cost-effective screening approach in the study.

Research Question 3

Of the three individual screening measures evaluated in this study, the DOR C_V across all grades was 0.46. When costs were factored in, ICER C_V across all grades was 0.54. These results suggest that performing a CEA does increase variability among screening measures. This result can be seen in practical terms when comparing the Forward Exam and the MAP. These measures are relatively similar in terms of DOR (Forward average DOR = 36.52, MAP average DOR = 32.59), but diverge in terms of ICER (Forward average ICER = $0.24, MAP average ICER = $0.61).

Research Question 4

The first sensitivity analysis increased high-level cost assumptions by 50%. In this sensitivity analysis, the average cost per student across grades for the Wisconsin Forward Exam increased from $7.72 to $8.05 while the average ICER increased from $0.238 to $0.248. For the MAP, the cost per student increased from $19.40 to $20.04; the average ICER increased from

$0.612 to $0.632. For MCAP, the average cost per student rose from $11.05 to $11.63 while the average ICER rose from $0.800 to $0.841. Similar modest increases in cost and ICER were found for both linear combinations of screening measures. For the Wisconsin Forward Exam and the MAP, a 50% increase in high-level assumption personnel costs results in an approximately 4% increase in ICER, while for MCAP the increase resulted in a 5% increase in ICER. Full results can be found in Table 3**.**

These results suggest that the CEAs conducted in this study are quite robust to variations in high-level cost assumptions. Indeed, for the Forward Exam to become a less cost-effective screening method than the MAP—the second most cost-effect single screening method evaluated—high-level cost assumptions would need to be increased by approximately 3,214%. To highlight the large degree of difference required to make the Wisconsin Forward Exam less cost-effective than the MAP, the time input of the SAC—for example—to prepare and administer the Forward Exam would need to increase from 15 hours to approximately 482 hours, or over 60 working days.

Table 4
Sensitivity Analyses by Screening Method and Grade

Method	Grade	Original Cost/Student	Original ICER	50% Increase Cost/Student	50% Increase ICER	50% Decrease Cost/Student	50% Decrease ICER
Forward							
	6	$7.57	$0.361	$8.29	$0.396	$6.84	$0.327
	7	$7.57	$0.182	$8.29	$0.199	$6.84	$0.164
	8	$8.01	$0.170	$8.74	$0.186	$7.29	$0.155
MAP							
	6	$19.40	$0.608	$20.04	$0.627	$18.77	$0.588
	7	$19.40	$0.739	$20.04	$0.763	$18.77	$0.715
	8	$19.40	$0.490	$20.04	$0.506	$18.77	$0.474
MCAP							
	6	$10.99	$0.692	$11.59	$0.728	$10.42	$0.656
	7	$11.08	$1.126	$11.65	$1.184	$10.51	$1.069
	8	$11.08	$0.582	$11.65	$0.611	$10.51	$0.552
Forward + MAP							
	6	$27.82	$0.691	$29.20	$0.715	$26.46	$0.657
	7	$27.82	$0.636	$29.20	$0.658	$26.46	$0.605
	8	$28.27	$0.423	$29.65	$0.438	$26.91	$0.403
Forward + MAP + MCAP							
	6	$38.81	$0.995	$40.79	$1.034	$36.88	$0.945
	7	$38.90	$0.910	$40.85	$0.646	$36.97	$0.865
	8	$39.35	$0.639	$41.29	$0.663	$37.42	$0.607

The second sensitivity analysis decreases high-level cost assumptions by 50%. In this analysis, the average cost per student across grades for the Wisconsin Forward Exam decreased from $7.72 to $6.99 and the average ICER decreased from $0.238 to $0.215. For the MAP, the cost per student decreased from $19.40 to $18.77 while the average ICER decreased from $0.612 to $0.592. For MCAP, the average cost per student fell from $11.05 to $10.48 while the average ICER fell from $0.800 to $0.759. As in the first sensitivity analysis, the pattern of small decreases was found in both multiple linear regression models as well. With a 50% decrease in high-level assumption personnel cost, Wisconsin Forward Exam average ICER decreased by 9.7%, MAP average ICER decreased by 3.3%, and MCAP average ICER decreased by 5.1%.

Practically speaking, decreases in high-level cost assumptions are even less impactful than cost increases as they are lower bound at zero. For example, even if high-level assumption costs for the MAP are decreased to nothing—where the MAP is prepared and administered without any school or district personnel other than a teacher proctoring, an almost impossible condition to imagine—the MAP is still far less cost-effective than the Wisconsin Forward Exam. Indeed, when these personnel costs are omitted entirely, the average ICER of the MAP only decreases from $0.612 to $0.572 (compared to an average ICER for the Forward Exam of $0.238).

CHAPTER FIVE

Discussion

The purpose of this dissertation is to help stakeholders make more informed decisions when selecting universal mathematics screening measures in middle school. There is a debate in the literature as to how to balance the need for measures that are accurate and for those that are efficient in terms of resources such as time and money. A corpus of research has been focused solely on the diagnostic accuracy of screening methods, but little attention has been paid to conducting rigorous studies on the costs of these methods. To help address these issues, I utilized CEA to determine which universal mathematics screening measures in middle school are most cost-effective, whether or not the use of CEA may provide more variability between screening methods to aid schools in selecting an approach, and to test the robustness of CEA results to changes in costs.

Summary of Results

Research question 1. The most cost-effective screening method examined in this study was the previous year's Forward Exam, beating out even linear combinations of multiple measures. The prior year Forward Exam had the highest diagnostic accuracy and the lowest cost per student of any individual screening method. The high diagnostic accuracy of the Forward Exam can reasonably be assumed to be the result of multiple factors. One of the simplest explanations is that the Forward Exam takes about twice as long to complete at the MAP, and over ten times as long to complete as a CBM. The MAP and CBM emphasize balancing accuracy administration time as they are intended for use as screening measures; the Forward Exam, as a criterion measure that is only given once yearly, is substantially longer and more

weighted heavily towards accuracy at the expense of a longer administration time. The administration format of the Wisconsin Forward Exam is only marginally changed year over year. Although the MAP and MCAP measure similar mathematics skills as the Forward Exam, the formats used to evaluate those skills are quite different. In short, it is reasonable to presume that, as the screening measure becomes more like the criterion measure in terms of format, diagnostic accuracy should increase. An additional consideration is student motivation. State-wide accountability testing has become an important benchmark and schools have significant incentives to perform at high levels. Schools and districts may be more likely to encourage students to give their best effort on state-wide testing as opposed to other screening methods such as MAP or CBM—which are administered multiple times per year.

The Forward Exam was also the least costly of any screening method analyzed. The explanation for this result is apparent in the data—the Forward Exam benefits greatly by not having an annual license fee for use. Licensing fees appear to be unique among the various screening ingredients; they must be paid on a per-student basis and greatly affect the overall cost per student of a screening measure. Licensing fees are not charged to districts as the State of Wisconsin incurs much of the cost of developing and administering this state-mandated test. Other costs, such as teacher time, can be spread evenly over many students, greatly reducing the cost of any individual ingredient per student. As an example, the Wisconsin Forward Exam requires a substantial input of time from the School Assessment Coordinator, usually a vice-principal or another administrator. In the CEA conducted, the cost of this one personnel ingredient was $1,127.87. But because the work of one SAC affects all students, this cost is divided by the number of students in a school, greatly reducing its impact (in this example, SAC time cost per student was $0.71). The cost of a license fee—for example, the $13.50 fee per

student to administer the MAP—cannot be spread over many students; although $13.50 is a small amount relative to the cost of a SAC, the cost recurs for every single student who takes the MAP.

The MAP was the second most cost-effective screening measure. It was easily the costliest screening method examined (largely due to annual licensing fees) but had relatively high diagnostic accuracy. In fact, the MAP was the most accurate single screener for Grade 6 students in this study. There do appear to be patterns that emerge when examining ICERs by grade. Screening costs tended and diagnostic accuracy tended to increase as a function of grade. Both these trends could be explained by increased administration time on the MCAP and the Forward Exam. As students aged, the MCAP CBM and Forward Exam consisted of more questions, leading to increased costs in terms of proctor time and likely increased diagnostic accuracy. Another factor that may be driving this trend was the actual increase in risk seen by grade. For all three screening measures, the proportion of true positives (the indication that students are not meeting proficiency) increased from Grade 6 to Grade 7, and from Grade 7 to Grade 8. It may be the case that risk is simply more pronounced in older grades, leading to increases in diagnostic accuracy. While interesting, any apparent trends emerging as a factor of age need to be considered with caution until replicated in other studies.

While the MAP was not as cost-effective as the Forward Exam, it may have additional benefits that were not considered in this study. Although a CEA is a useful and valid way to consider both costs and effects, it is limited in scope. A cost-benefit analysis (CBA) can incorporate advantages that are not measured in effect sizes. In the case of the MAP, one can readily identify possible benefits relative to the Forward Exam that are not quantified by diagnostic accuracy alone. Students move in between districts and States, meaning they may not

have the previous years' state-wide test data to be used as a screening method. The Forward Exam is only given once per year, making progressing monitoring difficult. The State of Wisconsin does not provide as many data utilization tools as many private companies such as the NWEA. What is it worth to have a screening measure that results in data for all students as opposed to most, is given thrice instead of once, or provides results that are easier to consume by school and district staff? These benefits are almost certainly worth something, but they must be evaluated using CBA.

Despite requiring the fewest resources to administer, the least cost-effective screening method examined was the MCAP. The MCAP resulted in very low diagnostic accuracy, below even acceptable standards. This lack of screening precision was not offset by MCAP's relatively low cost. Indeed, even if one eliminates the $6.50 license fee for MCAP and only factors in the cost of teacher proctoring and scoring, it still is not as cost-effective as the Wisconsin Forward Exam. Much like the MAP, however, MCAP may have additional benefits that could not be addressed using CEA. One consideration is the additional resources that come as a part of the MCAP licensing fee. MCAP is only one of the CBM that users gain access to as part of Pearson's AIMSweb testing tools. If administering multiple CBM in multiple subjects, the cost of the license fee would be spread over an increasing number of students, or, in another sense, a school psychologist could obtain much more data than MCAP alone could provide without paying additional fees. Assuming most school staff would choose to use multiple CBM and not MCAP alone, the cost per student of MCAP is likely overestimated in this study.

Another factor that might lead school personnel to use CBM for screening is immediacy. MCAP is the only screening method examined that can provide usable data within minutes of administration. This allows teachers and school psychologists to intervene quickly, potentially

ameliorating students' underperformance before they fall behind. CBM are, because of their short administration time and immediate results, also ideal for progress monitoring. While the Forward Exam and MAP give far more accurate predictions of future performance based on a single screening administration, neither are ideal for tracking student growth. CBM can be administered multiple times per week if needed and can provide useful data for making even high-stakes decisions such as special education qualification. The value of having a screening measure that can be administered repeatedly and whenever needed cannot be determined using CEA, but it almost certainly exists.

Research question 2. The most effective linear combination of screening approaches is Forward Exam + MAP. Although this combination was far more expensive than any individual measure, it also resulted in the highest diagnostic accuracy of any method or combination of methods. Of the 1,587 students who completed both the Forward Exam and the MAP, only 55 (3.5%) would be categorized as false negatives (the most problematic screening error) using this method. In comparison, for the Forward Exam by itself, the false-negative rate is over 11%. Although diagnostic accuracy is highest using this approach, the Wisconsin Forward Exam alone remains the most cost-effective approach overall due to its low cost. Despite coming in second in terms of cost-effectiveness, the linear combination of Forward + MAP should be strongly considered when school personnel select a screening approach. By using two sources of data, school psychologists can address the issue of students moving in and out of districts and states, can get some idea of the degree of regression students experienced over the summer, and can be sure that the fewest number of students who need early intervention will be missed during screening.

The linear combination of Forward + MAP + MCAP was the least cost-effective screening method examined in this study. A combination of exorbitant cost per student—over five times the cost of the Forward Exam alone—and decreased diagnostic accuracy resulting from the inclusion of MCAP in the model makes this screening approach one that should likely be avoided. Although this linear combination does result in higher diagnostic accuracy than any individual screening measure, the results of the CEA indicate that it would be more efficient for schools and districts to use a less costly screening approach and utilize the savings elsewhere. As an example, by choosing to use the Forward Exam only instead of the full linear combination, schools would save $31.30 per student (using the costs calculated for this study). When this is aggregated for all students who participated, it results in a cost-savings of $49,673.10—about 80% of the cost of hiring an early career teacher full-time with benefits.

This result highlights the utility of conducting CEA. If the goal of school personnel is to select the best screening measure without concern for costs, a simple comparison of diagnostic accuracy indices can be conducted. But for almost all schools, costs are a real factor of concern. By conducting a CEA, the school in question could select the most cost-effective screening approach (the previous year's Forward Exam) instead of a less cost-effective one (such as the linear combination of Forward + MAP + MCAP). In this scenario, there is only a small decrease in DOR (from average DOR = 47.795 to average DOR = 36.522), but many costs are avoided. Indeed, with the money saved, this school could pay the salary (without benefits) of a full-time early career teacher, purchase 25 new SMART Boards, or have 80% of the funds needed to construct—from scratch—a math intervention room that can serve four to five students at a time (Hollands et al., 2015).

Research question 3. One problem schools and districts encounter when selecting a screening approach from several screening measures may be a general lack of variance in diagnostic accuracy. As most researchers agree on minimum standards of sensitivity and specificity, a relatively high floor is set for screening measures to be considered valid for high-stakes decision-making. Because of this high floor, screening methods tend to cluster together in terms of diagnostic accuracy, making it difficult to determine which may be the most effective for a school or district. For example, using the data collected for this study, when local cut-scores are created so that every screening measure has SE ≥ .90, the SP of the MAP, Forward Exam, and the linear combination of both had SP values of .78, .79, and .83, respectively. With SP among the three methods so similar, it is easy to see how the selection of one approach can be difficult.

However, by combining both information on diagnostic accuracy and cost, the selection of one method becomes somewhat easier. The C_V of DOR between screening methods was .46, compared to .54 when using ICER. This result suggests that using CEA provides more information about the differences between screening measures, decreasing homogeneity and making the ultimate selection of one method over another simpler for stakeholders.

Research question 4. As this study used post-hoc estimates of most screening costs, it is reasonable to question whether the results for research questions 1 and 2 are robust to potential estimation decisions. Results demonstrate that the overall findings of this study—such as the fact that the Forward Exam was the most cost-effective single screening method—do not change even with large adjustments to ingredients costs. Despite increasing and decreasing high-level cost assumptions by 50%, ICERs remained remarkably stable. When costs estimates were

increased, ICERs increased by only $0.025 across all screening approaches. When cost estimates were decreased, ICERs fell by only $0.031 across all approaches.

If one attempted to imagine the most ideal conditions for the primary results of this study to be invalidated, it would include under-estimates of high-level assumption costs for some measures and over-estimates for others. If we assume this is the case (to test the robustness of the primary results), high-level cost assumptions for the MAP could be reduced to zero (resulting in an adjusted average ICER of $0.572) while costs for the Forward could be radically increased. In this example, even considering an administration of the MAP without any personnel other than a teacher proctor, the high-level cost assumptions of the Forward Exam would still need to be increased by 3,004% to make it less cost-effective than the MAP overall.

Taken together, the results of these two sensitivity analyses strongly support the robustness of the primary analysis results against even large changes in high-level cost assumptions. The principal reason for this is that these types of costs—where larger assumptions need to be made—make up a relatively small portion of costs overall. For example, the MAP licensing fee alone is more than high-level assumption personnel costs by a factor of 10. Although it is almost certainly the case that the high-level assumptions made about personnel costs in this study are not exactly accurate to real-world conditions of preparing and administering universal academic screening, they are almost certainly not erroneous enough to substantively impact the results of this study.

Exploratory results. One exploratory question that arose during this study was whether findings from the medical literature—that screening costs tend to outpace screening accuracy as more complex screening approaches are developed—could be found in these data. In the best-case scenario, where a school or district moves from using the best single screening method (the

Forward Exam) to the best linear combination of screening methods (Forward + MAP), DOR increases by 79.46% while cost per student increases by 119.83%. These percentage changes were calculated by averaging out DOR and cost per student of all three individual screening measures evaluated in the study and comparing them to the DOR and cost per student of Forward + MAP. These preliminary results suggest that, in general, administering multiple high-quality screening measures may increase costs more than it increases diagnostic accuracy. It should be noted that these findings must be interpreted with caution. The association between costs outpacing accuracy more complex screening approaches may only hold for the exact conditions of this study. Selecting different individual screening methods than those used in the study, combining those methods into different linear combinations, or increasing the number of measures included in a linear model could change drastically change the relationship between increased costs and increased DOR.

Results in the Context of Previous Research. Previous research has attempted to consider the costs of universal academic screening when examining diagnostic accuracy (VanDerHeyden et al., 2018; Klingbeil et al., 2019). VanDerHeyden and colleagues (2018) concluded that although the MAP was slightly more effective than CBM, the MAP was hindered by the high cost of administration. The present study supports part of the conclusion as results show that the MAP is over twice as costly as CBM. However, despite being more costly than CBM, MAP is more cost-effective due to its greater diagnostic accuracy. This shows the potential value of applying CEA to universal screening rather than trying to estimate the costs rudimentarily and comparing similar diagnostic accuracy metrics.

Klingbeil et al. (2019) listed screening approaches by cost category: (a) negligible, (b) minimal, (c) more costly, and (d) most costly. Importantly, Klingbeil and colleagues examined

the effect of creating local cut-scores (i.e., discrimination thresholds) which this study did not consider. Screening approaches were categorized based on what screening measures were already being implemented in the district. In other words, as the district was already administering both the Forward Exam and the MAP, the cost of those approaches were categorized "negligible" as they required no new inputs of time or money. In the "minimal cost category" Klingbeil et al. (2019) listed the 2016 Forward Exam (with local cut scores), the MAP (with local cut-scores) and the linear combination of the 2016 Forward Exam and the MAP. The authors suggested this categorization was due to the minimal amount of time required to estimate local cut-scores (using similar procedures to those used in the study). In the "more costly" category was MCAP because it required the administration of a novel measure that was not currently being administered by the district. Finally, the linear combination of the Forward Exam, MAP, and MCAP were in the "most costly" category as it required the collection of new data and additional statistical analyses to interpret. Notably, Klingbeil et al. (2019) did not provide an operational definition of these costs.

The results of this study, which calculated costs much more systematically, largely support these conclusions from Klingbeil and colleagues. The Forward Exam does indeed appear to be the least costly approach while any linear combination of approaches appears to be the most expensive. However, while Klingbeil et al. categorized MCAP as more costly, the results of this dissertation suggest that MCAP may be better categorized with the Forward Exam as minimally costly. It is important to note that, once diagnostic accuracy is factored in, MCAP appears to be an inferior screening method, despite being relatively inexpensive compared to other measures.

The results of this study differed from one suggestion from VanDerHeyden and Burns (2018). In their article, the authors stated that administering more than one screening measure is more costly and does not improve diagnostic accuracy. While this may indeed be the case in some instances, the results of the current study suggest that using multiple high-quality screening measures in conjunction (such as the Forward Exam and the MAP) can indeed increase diagnostic accuracy. And although administering multiple measures is almost certainly more costly than administering a single measure, that may not mean multiple measures are less *cost-effective*. In the study it was found that the combination of the 2016 Forward Exam and the MAP was more cost-effective than either the MCAP or MAP alone.

Results in the Context of Screening Costs. Two factors that may have an effect on conducting CEA on universal academic screening methods were not considered as a part of this study. First, many measures used for universal screening can be used for multiple purposes. For instance, the MAP is typically administered three times per year and can be used to help monitor the progress of students (and classrooms and schools) within and across school years. Similarly, MCAP can be used to measure student progress in response to intervention and as part of a comprehensive special education evaluation. Further, the cost of a license to use MCAP also includes the entire AIMSweb suite of CBMs which can be used to assess student skills in additional areas of reading and math. In short, the cost of many screening measures can be conceptualized in terms of utility. If a school or district is using the MAP for both screening and progress monitoring, the cost of the MAP can be adjusted downward when conducting a CEA on screening alone, thereby accounting for the fact that the tool has multiple purposes. Because the scope of this study was limited solely to academic screening and not, say, progress monitoring,

the costs (such as licensing fees) were not adjusted. The best method for estimating how these costs should be adjusted is an important unanswered question and a direction for future research.

Another factor that should be noted is the lack of a licensing fee for the Forward Exam. Licensing fees are essential to private businesses that must take in revenue to pay for the high cost of measure development. The Forward Exam, however, has no such fees as it is created, tested, and validated by the Wisconsin DPI. In this way, the Forward Exam is considered a public good, and the costs of developing and updating the measure are covered by Wisconsin taxpayers. This fact highlights the importance for practitioners to have a strong conceptualization of a CEA before it is performed. If this study were aimed at conducting a CEA on the Forward Exam from a State or Federal government standpoint, the cost of developing the Forward Exam should be included in the CEA as those funds could potentially be used on other government programs. The CEAs in this study were, however, not conducted from the standpoint of governments but of individual schools and districts. Although it most certainly costs money to develop and maintain the Forward Exam, those costs are not shared by schools through licensing fees or other revenue generation. As such, these costs were excluded from the analyses in this study.

Limitations

Several limitations must be acknowledged for this study. The largest factor impacting the external validity of these results are the screening measures tested. The results of this study will only hold when using the same screening measures in a similar district. Using a different criterion measure, different screening approaches, different time points (such as winter for the MAP instead of fall), or even modes of administration (many CBM are not computerized, for instance) would all very likely affect the results. As more research in this area is conducted, it

may be possible to identify trends in the cost-effectiveness of various screening approaches that can be used to create general principles. But as this is the first study to use CEA in universal academic screening, it will remain important for schools and districts to conduct their own CEA when attempting to determine the most cost-effective screening approach.

A second limitation is the calculation of ingredient costs. Despite sensitivity analyses that demonstrate the robustness of results to changes in high-level assumptions, there are simply too many variables to consider when conducting a post-hoc CEA. For example, when calculating the cost for the Data Administrator ingredient of the MAP, I choose to assume that the process of enrolling students into the MAP roster would be from scratch. While this is a reasonable assumption for a school or district that has never administered the MAP, it is not for schools and districts that regularly administer the screener. Once the initial roster is built, it only needs to be updated with students matriculating into or out of a school or district, a time commitment that should be much less than building a roster from scratch. Although this decision did not impact the main findings (it was shown in the sensitivity analyses that one could eliminate all personnel ingredients other than a proctor and the MAP would still not be as cost-effective as the Forward Exam), it certainly affects the derived ICERs. In future research, every effort should be made to collect exact costs during screening preparation and administration instead of post-hoc assumptions.

A third limitation is the inability of CEA to incorporate benefits not measured in effect sizes. Conducting a CEA on mathematics screening measures suggests that the only metrics of interest are cost and diagnostic accuracy. There are many other factors that schools and districts may consider when selecting a screening approach such as the delay between test administration and data analysis, the usability of data to inform decision making, the degree to which student

progress can be monitored, or the usefulness of the screening measures to inform other decisions required in MTSS frameworks. The results of this study suggest that using the previous year's Forward Exam is the most cost-effective screening method for schools to use to predict math risk in middle school. However, this should not be interpreted as prescriptive. Schools and districts with highly mobile student populations, for instance, may decide rightly that another screening approach is more desirable as it has the benefit of producing data for all students, regardless of how long they have been in a district.

Fourth, equity is an issue that was not addressed in this study but should be considered by all researchers and practitioners. A CEA should be considered as one tool for selecting mathematics screening measures. Universal screening is an essential component of MTSS, but MTSS itself can rest on spurious assumptions for schools and districts with large proportions of marginalized students. The MTSS model in general, and the RTI model in particular, work on the assumption that only a small percentage of students in a school will require academic intervention. In schools struggling to contend with the opportunity gap, this is seldom the case. Indeed, in a large urban district that neighbors the district that participated in this study, only approximately 16% of students reached proficiency in mathematics on the 2018 Forward Exam. In such districts, the utility of any universal screening at all is questioned; if over 80% of students likely require some type of tier II or tier III mathematics intervention, it may be more cost-effective to eliminate screening entirely and dedicate those resources to interventions delivered at tier I.

A final limitation concerns my assumption of perfect utilization when calculating ingredient costs. The costs for ingredients such as laptops or teacher time are calculated by dividing the total cost per year of the ingredient by the amount of time that it is required to

administer a screening measure. In the case of both teacher time and laptops, this method for calculating costs assumes perfect utilization of ingredients outside of screening. If a laptop were in use for any academic activity for every minute of every school day, it would be perfectly utilized and the cost of use per minute can be easily calculated (as it was in this study). It is unreasonable, however, to expect this to be the case. Laptops require charging, are not used in every class or for every project, etc. When ingredients are not perfectly utilized, they become more expensive. If a laptop, for instance, was bought by a district but only ever used for MAP testing, then the entire cost of the laptop would have to be used in a CEA on math screening. Although the idea of purchasing a laptop for use two or three times a year is as farfetched as expecting perfect utilization, it is almost certainly the case that the costs of many ingredients were inflated in this study due to imperfect utilization.

Utilization is also a concern when considering licensing fees. For both the MAP and CBM, the annual license fee was included as an ingredient cost. However, for both measures, the license fee provides more than just a screening measure for mathematics. As discussed above, both the MAP and MCAP CBM are used for progress monitoring as well. Ideally, researchers would determine how often MAP data (for instance) were used as screening data in a school year, and how often they were used as progress monitoring data. The license fee ingredient would then be adjusted to reflect this split.

Implications and Future Directions

This study provides several findings that demonstrate the utility of CEA on universal mathematics screening in middle schools. By conducting CEA, variability among screening methods was increased making selection easier, ICERs were calculated for three screening measures and two combinations of measures, and the cost-effectiveness of combining certain

screening approaches was supported. When costs are measured as a normal practice of administering academic screening in mathematics, conducting a CEA appears to be a relatively inexpensive and useful tool to help schools choose a screening approach.

While broad recommendations are not appropriate given the lack of corroborating research in this area, some preliminary suggestions for practice can be made. Schools and districts will likely find benefits in regularly conducting CEA to help make difficult decisions such as determining an approach to universal screening in mathematics. Using the previous year's criterion measure as a screener for the following year appears to be the most cost-effective approach to universal middle school screening in mathematics and should be considered for more widespread adoption. Additional screening measures such as the MAP should likely only be considered for usage if they meet minimum standards of diagnostic accuracy. When multiple high-quality screeners are administered, it appears to be a good use of time and resources to have school psychologists combine these data using multiple linear regression. Finally, using CEA may help schools and districts serve their students more effectively by making more efficient use of limited resources.

One consideration that schools and districts will have to contend with is the heterogeneous nature of screening approaches and criterion measures across the nation. While this study examined MCAP, MAP, and the Forward Exam as they are widely used (or required) in Wisconsin, other States use different criterion measures and may use different screening approaches. Despite this fact, schools and districts can apply the method for conducting CEA that was employed in this study, regardless of the actual measures used. For schools that do not use the MAP or CBM for screening purposes, stakeholders should first determine which screening measures are currently being used and which screening measures may be likely to be

adopted in the future. Practitioners may then pilot said screening measures to obtain local diagnostic accuracy data or utilize data provided by the screening vendor or peer-reviewed research. In addition, the Every Student Succeeds Act requires that all states administer an annual statewide test of reading/language arts and math. The methods used to evaluate the Forward Exam in this study can be applied to other similar annual statewide tests.

The National Center on Intensive Intervention (NCII), developed by the American Institutes for Research (and supported by the United States Department of Education) compiles a database of screening measures used all over the country (NCII, 2020). Each measure is categorized by classification accuracy, technical standards, and usability and can be a useful tool for stakeholders when attempting to determine which screening measures might be adopted by schools or districts. After one or more screening measures are selected for a CEA, schools should use their individual end-of-year statewide test as a criterion measure to evaluate the diagnostic accuracy of the screening approaches selected. In this fashion, schools and districts can conduct their own CEA on universal academic screening using local measures as opposed to those selected for analysis in this study.

One potential future direction of CEA research in universal academic screening is the possible inclusion of extant data in screening. At this point, universal academic screening does not factor in student-level data that could be used to make more accurate screening decisions. A student's risk status in previous academic years, socio-economic status, or attendance rates may provide additional useful information in predicting student performance. It is possible that these data could be used in conjunction with academic screening data to increase diagnostic accuracy without increasing costs. There are, however, some drawbacks to this approach.

Socio-economic status or attendance data may be correlated with other student-level characteristics such as ethnicity. The inclusion of these types of data into a universal academic screening approach may result in unintended segregation of students. For example, if two students attain the same score on the MAP, but one student qualifies for free-and-reduced lunch (FRL) whereas the other does not (and these data were included with MAP data in a linear or logistic regression), the student who qualifies for FRL could be classified as at-risk whereas the student who does not qualify could be classified as not-at-risk. Although this may indeed be the correct decision at the student level, it is not hard to imagine how this approach at the school, district, or State level may lead to widespread tracking that is, at least in part, based on non-academic factors such as family income.

Further research is required to both validate and extend the results of this study. Additional studies examining many different screening methods and criterion measures can help extend the implications of this study to other districts and other States that do not commonly use the MAP or Forward Exam. Research conducting CEAs with exact ingredient costs instead of post-hoc cost estimations should add additional legitimacy to the field of study. Researchers can extend the findings of this study by conducting cost-benefit analyses to account for factors that cannot be readily quantified in CEA. Finally, replicating this study in a district with a larger proportion of marginalized students and families can help verify that conducting CEA is an equitable practice.

Conclusion

Researchers and practitioners have debated which universal academic mathematics screening measures are best for utilization in schools for some time. Widespread adoption of the MTSS model requires some form of academic screening to ensure that students requiring more

intensive intervention are identified. Metrics of diagnostic accuracy are traditionally used to select screening approaches, but many screening measures cluster on this metric. A CEA was conducted that demonstrated increased variability among measures, identified the most cost-effective approaches, and validated the practice of combining multiple measures using multiple linear regression.

This study provides initial evidence of the utility of conducting a CEA to evaluate universal academic screening. Schools and districts could follow similar steps to estimate the diagnostic accuracy of their screening procedures and estimate the associated costs to conduct empirically valid CEA in order to help make decisions about universal academic screening. It has corroborated previous research in showing that are indeed substantial cost differences in screening approaches and expanded on that research by combining data on cost and diagnostic accuracy. The results of this study appear to be very robust to errors in the estimation of costs, further providing evidence that stakeholders can conduct CEA prior to actually implementing any new screening approach. These results demonstrate the need for further research into this emerging field and provide some initial guidance for practitioners attempting to select screening measures.

References

Abdi, H. (2010). Coefficient of Variation. In N. Salkind (Ed.), *Encyclopedia of Research Design* (Vol. 1, pp. 169–171). SAGE Publications Inc.

Albers, C. A., Glover, T. A., & Kratochwill, T. R. (2007). Where are we, and where do we go now? *Journal of School Psychology, 45*(2), 257–263. https://doi.org/10.1016/j.jsp.2006.12.003

Arbel, R., & Greenberg, D. (2016). Rethinking cost-effectiveness in the era of zero healthcare spending growth. *International Journal for Equity in Health*, *15*(1), 33. https://doi.org/10.1186/s12939-016-0326-8

Bahr, M. W., Leduc, J. D., Hild, M. A., Davis, S. E., Summers, J. K., & McNeal, B. (2017). Evidence for the Expanding Role of Consultation in the Practice of School Psychologists: Evidence for the Expanding Role. *Psychology in the Schools*, *54*(6), 581–595. https://doi.org/10.1002/pits.22020

Ball, C. R., & Christ, T. J. (2012). Supporting valid decision making: Uses and misuses of assessment data within the context of RTI: Supporting Valid Decision Making. *Psychology in the Schools*, *49*(3), 231–244. https://doi.org/10.1002/pits.21592

Barrett, C. A., Gadke, D. L., & VanDerHeyden, A. M. (2020). At What Cost?: Introduction to the Special Issue "Return on Investment for Academic and Behavioral Assessment and Intervention." *School Psychology Review, 49*(4), 347–358. https://doi.org/10.1080/2372966X.2020.1817718

Barrett, C. A., & VanDerHeyden, A. M. (2020). A cost-effectiveness analysis of classwide math intervention. *Journal of School Psychology*, *80*, 54–65. https://doi.org/10.1016/j.jsp.2020.04.002

Burns, M. K., Codding, R. S., Boice, C. H., & Lukito, G. (2010). Meta-analysis of acquisition and fluency math interventions with instructional and frustration level skills: Evidence for a skill-by-treatment interaction. *School Psychology Review, 39,* 69-83.

Catts, H. W., Petscher, Y., Schatschneider, C., Sittner Bridges, M., & Mendoza, K. (2009). Floor Effects Associated With Universal Screening and Their Impact on the Early Identification of Reading Disabilities. *Journal of Learning Disabilities, 42*(2), 163–176. https://doi.org/10.1177/0022219408326219

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the Risk of Future Reading Difficulties in Kindergarten Children: A Research-Based Model and Its Clinical Implementation. *Language, Speech, and Hearing Services in Schools*, *32*(1), 38–50. https://doi.org/10.1044/0161-1461(2001/004)

Cheung, A. C. K., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review, 9*, 88–113. https://doi.org/10.1016/j.edurev.2013.01.001

Chodura, S., Kuhn, J.-T., & Holling, H. (2015). Interventions for Children With Mathematical Difficulties: A Meta-Analysis. *Zeitschrift Für Psychologie, 223*(2), 129–144. https://doi.org/10.1027/2151-2604/a000211

Christ, T. J., & Nelson, P. M. (2014). Developing and evaluating screening systems: 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). American Psychological Association. https://doi.org/10.1037/14316-004

Christ, T. J., Nelson, P. M., Van Norman, E. R., Chafouleas, S. M., & Riley-Tillman, T. C. (2014). Direct Behavior Rating: An evaluation of time-series interpretations as consequential validity. *School Psychology Quarterly*, *29*(2), 157–170. https://doi.org/10.1037/spq0000029

Clemens, N. H., Keller-Margulis, M. A., Scholten, T., & Yoon, M. (2016). Screening Assessment Within a Multi-Tiered System of Support: Current Practices, Advances, and Next Steps. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of Response to Intervention: The Science and Practice of Multi-Tiered Systems of Support* (pp. 187–213). Springer US. https://doi.org/10.1007/978-1-4899-7568-3_12

Cohen, D. J., & Reynolds, M. R. (2008). Interpreting the Results of Cost-Effectiveness Studies. *Journal of the American College of Cardiology*, *52*(25), 2119–2126. https://doi.org/10.1016/j.jacc.2008.09.018

Cope, B., & Kalantzis, M. (2016). Big Data Comes to School: Implications for Learning, Assessment, and Research. *AERA Open*, *2*(2), 233285841664190. https://doi.org/10.1177/2332858416641907

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

Elliott, S. N., Huai, N., & Roach, A. T. (2007). Universal and early screening for educational difficulties: Current and future approaches. *Journal of School Psychology*, *45*(2), 137–161. https://doi.org/10.1016/j.jsp.2006.11.002

Forman, S. G., Olin, S. S., Hoagwood, K. E., Crowe, M., & Saka, N. (2009). Evidence-Based Interventions in Schools: Developers' Views of Implementation Barriers and Facilitators. *School Mental Health*, *1*(1), 26–36. https://doi.org/10.1007/s12310-008-9002-5

Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly*, *41*(1), 93–99. https://doi.org/10.1598/RRQ.41.1.4

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

Fuchs, L. S., Fuchs, D., & Zumeta, R. O. (2008). A Curricular-Sampling Approach to Progress Monitoring: Mathematics Concepts and Applications. *Assessment for Effective Intervention*, *33*(4), 225–233. https://doi.org/10.1177/1534508407313484

Gersten, R., Clarke, B., Jordan, N. C., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal Screening in Mathematics for the Primary Grades: Beginnings of a Research Base. *Exceptional Children*, *78*(4), 423–445. https://doi.org/10.1177/001440291207800403

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology*, *45*(2), 117–135. https://doi.org/10.1016/j.jsp.2006.05.005

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. Annual Review of Psychology, 60(1), 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, *4*(2), 627–635.

Harbison, R. W., & Hanushek, E. A. (1992). Educational performance for the poor: Lessons from rural northeast Brazil. Oxford, UK: Oxford University Press.

Hollands, F., Bowden, A. B., Belfield, C., Levin, H. M., Cheng, H., Shand, R., Pan, Y., & Hanisch-Cerda, B. (2014). Cost-Effectiveness Analysis in Practice: Interventions to Improve High School Completion. *Educational Evaluation and Policy Analysis, 36*(3), 307–326. https://doi.org/10.3102/0162373713511850

Hollands, F.M., Hanisch-Cerda, B., Levin, H. M., Belfield, C.R., Menon, A., Shand, R., Pan, Y., Bakir, I., & Cheng, H. (2015). CostOut - the CBCSE Cost Tool Kit. Center for Benefit-Cost Studies of Education, Teachers College, Columbia University. Retrieved from: www.cbcsecosttoolkit.org

Hummel-Rossi, B., & Ashdown, J. (2002). The State of Cost-Benefit and Cost-Effectiveness Analyses in Education. *Review of Educational Research, 72*(1), 1–30. https://doi.org/10.3102/00346543072001001

Hunter, L. J., DiPerna, J. C., Hart, S. C., & Crowley, M. (2018). At what cost? Examining the cost effectiveness of a universal social–emotional learning program. *School Psychology Quarterly, 33*(1), 147–154. https://doi.org/10.1037/spq0000232

Individuals with Disabilities Education Improvement Act (2004). Pub. L. No. 108-466.

January, S.-A. 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. https://doi.org/10.1177/1534508415579095

Johnson, E. S., Jenkins, J. R., & Petscher, Y. (2010). Improving the Accuracy of a Direct Route Screening Process. *Assessment for Effective Intervention, 35*(3), 131–140. https://doi.org/10.1177/1534508409348375

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How Can We Improve the Accuracy of Screening Instruments? *Learning Disabilities Research & Practice, 24*(4), 174–185. https://doi.org/10.1111/j.1540-5826.2009.00291.x

Keren, R., Helfand, M., Homer, C., McPhillips, H., & Lieu, T. A. (2002). Projected Cost-Effectiveness of Statewide Universal Newborn Hearing Screening. *PEDIATRICS*, *110*(5), 855–864. https://doi.org/10.1542/peds.110.5.855

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

Klingbeil, D. A., Maurice, S. A., Van Norman, E. R., Nelson, P. M., Birr, C., Hanrahan, A. R., Schramm, A. L., Copek, R. A., Carse, S. A., Koppel, R. A., & Lopez, A. L. (2019). Improving Mathematics Screening in Middle School. *School Psychology Review*, *48*(4), 383–398. https://doi.org/10.17105/SPR-2018-0084.V48-4

Klingbeil, D. A., Nelson, P. M., Van Norman, E. R., & Birr, C. (2017). Diagnostic Accuracy of Multivariate Universal Screening Procedures for Reading in Upper Elementary Grades. *Remedial and Special Education*, *38*(5), 308–320. https://doi.org/10.1177/0741932517697446

Klingbeil, D. A., Van Norman, E. R., Nelson, P. M., & Birr, C. (2018). Evaluating Screening Procedures Across Changes to the Statewide Achievement Test. *Assessment for Effective Intervention*, *44*(1), 17–31. https://doi.org/10.1177/1534508417747390

Levin, H. (1975). Cost-effectiveness analysis in evaluation research. In M. Guttentag & E. L. Struening (Eds.), Handbook of evaluation research (Vol. 2; pp. 346–368). Beverly Hills, CA: Sage.

Levin, H. (2001). Waiting for Godot: Cost-Effectiveness Analysis in Education. *New Directions for Evaluation*, *2001*(90), 55. https://doi.org/10.1002/ev.12

Levin, H. M. (1991). Raising Productivity in Higher Education. *The Journal of Higher Education*, *62*(3), 241. https://doi.org/10.2307/1982281

Levin, H. M., & Belfield, C. (2015). Guiding the Development and Use of Cost-Effectiveness Analysis in Education. *Journal of Research on Educational Effectiveness*, *8*(3), 400–418. https://doi.org/10.1080/19345747.2014.915604

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications* (2nd ed). Sage Publications.

Little, R. J. A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

McIntosh, K., & Goodman, S. (2016). *Integrated multi-tiered systems of support: Blending RTI and PBIS*. New York, NY: Guilford Publications.

Mitchell, A. S. (2002). Anitpodean Assessment: Activities, Actions, and Achievements. *International Journal of Technology Assessment in Health Care*, *18*(2), 203–212. https://doi.org/10.1017/S0266462302000223

National Education Association (2018). 2017-2018 Average Starting Teacher Salaries by State. (n.d.). Retrieved May 28, 2019, from http://www.nea.org/home/2017-2018-average-starting-teacher-salary.html

National Center for Education Statistics (2020). *Mathematics Performance*. Retrieved from https://nces.ed.gov/programs/coe/indicator_cnc.asp.

National Center for Education Statistics (2012a). 2017-2018 Average Starting Teacher Salaries by State. (n.d.). Retrieved May 28, 2019, from http://www.nea.org/home/2017-2018-average-starting-teacher-salary.html

National Center for Education Statistics (2012b). Schools and Staffing Survey (SASS). (n.d.). Retrieved April 30, 2019, from https://nces.ed.gov/surveys/sass/tables/sass1112_2013314_t1s_007.asp

National Center for Intensive Intervention (2020). Academic Screening Tools Chart. (n.d.). Retrieved July 24, 2021, from https://charts.intensiveintervention.org/ascreening

Nelson, P. M., Van Norman, E. R., & Lackner, S. K. (2016). A Comparison of Methods to Screen Middle School Students for Reading and Math Difficulties. *School Psychology Review, 45*(3), 327–342. https://doi.org/10.17105/SPR45-3.327-342

Neumann, P. J. (2004). Why don't Americans use cost-effectiveness analysis? *The American Journal of Managed Care*, *10*(5), 308–312.

Northwest Evaluation Association (2011). *Technical Manual for Measures of Academic Progress (MAP) and Measures of Academic Progress or Primary Grades (MPG).* Portland, OR: Author.

Northwest Evaluation Association (2017). *Linking the Wisconsin Forward Exam to NWEA MAP Tests.* Retrieved from nwea.org.

Northwest Evaluation Association (2018). *2018-2019 MAP Test Administration Manual.* Retrieved from nwea.org.

Northwest Evaluation Association (2020a). *Assessment Coordination Guide*. Retrieved from nwea.org.

Northwest Evaluation Association (2020b). *Students and Staff Management Guide*. Retrieved from nwea.org.

Northwest Evaluation Association (2020c). *Proctor Guide*. Retrieved from nwea.org.

NCS Pearson, Inc. (2012). *AIMSweb Technical Manual.* Bloomington, MN: Author

NCS Pearson, Inc. (2017). *aimswebPLUS Technical Manual*. Bloomington, MN: Author

Paulden, M. (2020). Calculating and Interpreting ICERs and Net Benefit. *PharmacoEconomics*, *38*(8), 785–807. https://doi.org/10.1007/s40273-020-00914-6

Okamoto, Y., & Case, R. (1996). II. EXPLORING THE MICROSTRUCTURE OF CHILDREN'S CENTRAL CONCEPTUAL STRUCTURES IN THE DOMAIN OF NUMBER. *Monographs of the Society for Research in Child Development*, *61*(1–2), 27–58. https://doi.org/10.1111/j.1540-5834.1996.tb00536.x

Peugh, J. L., & Enders, C. K. (2004). Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research*, *74*(4), 525–556. https://doi.org/10.3102/00346543074004525

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. *Learning Disabilities Research & Practice*, *27*(3), 136–147. https://doi.org/10.1111/j.1540-5826.2012.00359.x

Quinn, B., Van Mondfrans, A., & Worthen, B. R. (1984). Cost-Effectiveness of Two Math Programs as Moderated by Pupil SES. *Educational Evaluation and Policy Analysis*, *6*(1), 39–52. https://doi.org/10.3102/01623737006001039

Salinger, R. L. (2016). Selecting Universal Screening Measures to Identify Students at Risk Academically. *Intervention in School and Clinic*, *52*(2), 77–84. https://doi.org/10.1177/1053451216636027

Shapiro, E. S., & Gebhardt, S. N. (2012). Comparing Computer-Adaptive and Curriculum-Based Measurement Methods of Assessment. *School Psychology Review*, *41*(3), 295–305. https://doi.org/10.1080/02796015.2012.12087510

Stormont, M., Herman, K. C., Reinke, W. M., King, K. R., & Owens, S. (2015). The Kindergarten Academic and Behavior Readiness Screener: The utility of single-item teacher ratings of kindergarten readiness. *School Psychology Quarterly*, *30*(2), 212–228. https://doi.org/10.1037/spq0000089

Strait, G. G., Smith, B. H., & McQuillin, S. D. (2018). Aggregated Randomly Generated Math Curriculum–Based Measurements for Middle School Students: Reliability, Predictive Validity, and Cut Score Precision. *Assessment for Effective Intervention*, *44*(1), 58–64. https://doi.org/10.1177/1534508418761231

Stoiber, K. C., & Gettinger, M. (2016). Multi-tiered systems of support and evidence-based practices. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), Handbook of response to intervention: The science and practice of multi-tiered systems of support (2nd ed., pp. 121–141). Springer.

Van Norman, E. R., Nelson, P. M., & Klingbeil, D. A. (2017). Single measure and gated screening approaches for identifying students at-risk for academic problems: Implications for sensitivity and specificity. *School Psychology Quarterly*, *32*(3), 405–413. https://doi.org/10.1037/spq0000177

Van Norman, E. R., Nelson, P. M., & Parker, D. C. (2017). Technical adequacy of growth estimates from a computer adaptive test: Implications for progress monitoring. *School Psychology Quarterly*, *32*(3), 379–391. https://doi.org/10.1037/spq0000175

Vanderheyden, A. M. (2011). Technical Adequacy of Response to Intervention Decisions. *Exceptional Children*, *77*(3), 335–350. https://doi.org/10.1177/001440291107700305

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

VanDerHeyden, A. M., & Burns, M. K. (2018). Improving Decision Making in School Psychology: Making a Difference in the Lives of Students, Not Just a Prediction About Their Lives. *School Psychology Review*, *47*(4), 385–395. https://doi.org/10.17105/SPR-2018-0042.V47-4

VanDerHeyden, A. M., Burns, M. K., & Bonifay, W. (2018). Is More Screening Better? The Relationship Between Frequent Screening, Accurate Decisions, and Reading Proficiency. *School Psychology Review*, *47*(1), 62–82. https://doi.org/10.17105/SPR-2017-0017.V47-1

VanDerHeyden, A. M., Codding, R. S., & Martin, R. (2017). Relative Value of Common Screening Measures in Mathematics. *School Psychology Review*, *46*(1), 65–87. https://doi.org/10.17105/SPR46-1.65-87

Vaughn, S., & Fletcher, J. M. (2012). Response to Intervention With Secondary School Students With Reading Difficulties. *Journal of Learning Disabilities*, *45*(3), 244–256. https://doi.org/10.1177/0022219412442157

Walker, H. M., & Shinn, M. R. (2002). Structuring school-based interventions to achieve integrated primary, secondary, and tertiary prevention goals for safe and effective

schools. In M. R. Shinn, H. M. Walker, & G.Stoner (Eds.), *Interventions for academic and behavior problems II: Preventative and remedial approaches* (pp. 1−25). Bethesda, MD: NASP.

Wisconsin Department of Public Instruction (2020a) *Test Administration Manual.* Retrieved from dpi.wi.gov.

Wisconsin Department of Public Instruction (2020b) *DRC Insight Portal Guide.* Retrieved from dpi.wi.gov.

Wisconsin Department of Public Instruction (2020c) *Site Technology Readiness Checklist for Deploying DRC Insight Online Assessments.* Retrieved from dpi.wi.gov.

Wisconsin Department of Public Instruction (2020d) *District Assessment Coordinator Checklist.* Retrieved from dpi.wi.gov.

Wisconsin Department of Public Instruction (2020e) *School Assessment Coordinator Checklist.* Retrieved from dpi.wi.gov.

Wisconsin Department of Public Instruction (2020f) *Test Administrator/Proctor Checklist.* Retrieved from dpi.wi.gov.

Wisconsin Department of Public Instruction (2016). *Wisconsin Forward Exam: Spring 2016 Technical Report.* Madison, WI: Author. Retrieved from: http://dpi.wi.gov

Wisconsin Department of Public Instruction (2019). *Forward Exam 2019 Suggested Test Times.* Madison, WI: Author. Retrieved from: http://dpi.wi.gov

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart1. *Journal of Applied Behavior Analysis, 11*(2), 203–214. https://doi.org/10.1901/jaba.1978.11-203

Yeo, S. (2010). Predicting Performance on State Achievement Tests Using Curriculum-Based Measurement in Reading: A Multilevel Meta-Analysis. *Remedial and Special Education*, *31*(6), 412–422. https://doi.org/10.1177/0741932508327463

APPENDIX

In this appendix, the reader will find tables showing ingredient costs in more detail. Tables A.1 through A.10 present individual ingredient costs for each individual and linear combination of screening measures, stratified by grade when costs change as a function of grade. Table A.11 shows the cost of personnel ingredients by screening method. Tables A.12 through A.14 present personnel costs as a function of price, quantity needed (in minutes), and the number of students served.

Table A.1
Ingredients Required for the 2016 Forward Exam Grade 6 & 7

Category	Ingredient	Cost/Student
Personnel	Classroom Teacher (training)	$1.309
	Classroom Teacher (administration)	$4.581
	District Assessment Coordinator (training)	$0.005
	District Assessment Coordinator (preparation and administration)	$0.258
	School Assessment Coordinator (training)	$0.036
	School Assessment Coordinator (preparation and administration)	$0.711
	District Technology Coordinator (training)	$0.014
	District Technology Coordinator (preparation and administration)	$0.193
	School Technology Coordinator (training)	$0.044
	School Technology Coordinator (preparation and administration)	$0.292
Facilities	N/A	
Materials	Computer (administration)	$0.124
Other	N/A	
Grand Total		$7.566

Table A.2
Ingredients Required for the 2016 Forward Exam Grade 8

Category	Ingredient	Cost/Student
Personnel	Classroom Teacher (training)	$1.309
	Classroom Teacher (administration)	$5.017
	District Assessment Coordinator (training)	$0.005
	District Assessment Coordinator (preparation and administration)	$0.258
	School Assessment Coordinator (training)	$0.036
	School Assessment Coordinator (preparation and administration)	$0.711
	District Technology Coordinator (training)	$0.014
	District Technology Coordinator (preparation and administration)	$0.193
	School Technology Coordinator (training)	$0.044
	School Technology Coordinator (preparation and administration)	$0.292
Facilities	N/A	
Materials	Computer (administration)	$0.136
Other	N/A	
Grand Total		$8.014

Table A.4
Ingredients Required for the MCAP Grade 6

Category	Ingredient	Cost/Student
Personnel	Classroom Teacher (training)	$2.618
	Classroom Teacher (administration)	$0.349
	Classroom Teacher (scoring)	$1.134
Facilities	N/A	
Materials	Annual License	$6.500
	Paper (administration)	$0.390
Other	N/A	
Grand Total		$10.991

Table A.5
Ingredients Required for the MCAP Grade 7 & 8

Category	Ingredient	Cost/Student
Personnel	Classroom Teacher (training)	$2.618
	Classroom Teacher (administration)	$0.436
	Classroom Teacher (scoring)	$1.134
Facilities	N/A	
Materials	Annual License	$6.50
	Paper (administration)	$0.390
Other	N/A	
Grand Total		$11.078

Table A.6

Ingredients Required for the 2016 Forward Exam + MAP Grade 6 & 7

Category	Measure	Ingredient	Cost/Student
Personnel			
	Forward	Classroom Teacher (training)	$1.309
		Classroom Teacher (administration)	$4.581
		District Assessment Coordinator (training)	$0.005
		District Assessment Coordinator (preparation and administration)	$0.258
		School Assessment Coordinator (training)	$0.036
		School Assessment Coordinator (preparation and administration)	$0.711
		District Technology Coordinator (training)	$0.014
		District Technology Coordinator (preparation and administration)	$0.193
		School Technology Coordinator (training)	$0.044
		School Technology Coordinator (preparation and administration)	$0.292
	MAP	Classroom Teacher (training)	$2.618
		Classroom Teacher (administration)	$1.963
		District Assessment Coordinator (preparation and administration)	$0.258
		School Assessment Coordinator (preparation and administration)	$0.711
		Systems Administrator (preparation and administration)	$0.061
		Data Administrator (preparation)	$0.239
Facilities		N/A	
Materials	Forward	Computer (administration)	$0.124
	MAP	Annual License	$13.50
		Computer (administration)	$0.053
Other		School Psychologist (data analysis)	$0.037
		SPSS Annual License	$0.813
Grand Total			$27.819

Table A.7

Ingredients Required for the 2016 Forward Exam + MAP Grade 8

Category	Measure	Ingredient	Cost/Student
Personnel			
	Forward	Classroom Teacher (training)	$1.309
		Classroom Teacher (administration)	$5.017
		District Assessment Coordinator (training)	$0.005
		District Assessment Coordinator (preparation and administration)	$0.258
		School Assessment Coordinator (training)	$0.036
		School Assessment Coordinator (preparation and administration)	$0.711
		District Technology Coordinator (training)	$0.014
		District Technology Coordinator (preparation and administration)	$0.193
		School Technology Coordinator (training)	$0.044
		School Technology Coordinator (preparation and administration)	$0.292
	MAP	Classroom Teacher (training)	$2.618
		Classroom Teacher (administration)	$1.963
		District Assessment Coordinator (preparation and administration)	$0.258
		School Assessment Coordinator (preparation and administration)	$0.711
		Systems Administrator (preparation and administration)	$0.061
		Data Administrator (preparation)	$0.239
Facilities		N/A	
Materials	Forward	Computer (administration)	$0.136
	MAP	Annual License	$13.50
		Computer (administration)	$0.053
Other		School Psychologist (data analysis)	$0.037
		SPSS Annual License	$0.813
Grand Total			$28.267

Table A.8

Ingredients Required for the 2016 Forward Exam + MAP + MCAP Grade 6

Category	Measure	Ingredient	Cost/Student
Personnel			
	Forward	All Personnel Costs	$7.442
	MAP	All Personnel Costs	$5.850
	MCAP	All Personnel Costs	$4.101
Facilities		N/A	
Materials	Forward	All Materials Costs	$0.124
	MAP	All Materials Costs	$13.553
	MCAP	All Materials Costs	$6.890
Other		School Psychologist (data analysis)	$0.037
		SPSS Annual License	$0.813
Grand Total			$38.810

Table A.9

Ingredients Required for the 2016 Forward Exam + MAP + MCAP Grade 7

Category	Measure	Ingredient	Cost/Student
Personnel			
	Forward	All Personnel Costs	$7.442
	MAP	All Personnel Costs	$5.850
	MCAP	All Personnel Costs	$4.188
Facilities		N/A	
Materials	Forward	All Materials Costs	$0.124
	MAP	All Materials Costs	$13.553
	MCAP	All Materials Costs	$6.890
Other		School Psychologist (data analysis)	$0.037
		SPSS Annual License	$0.813
Grand Total			$38.897

Table A.10

Ingredients Required for the 2016 Forward Exam + MAP + MCAP Grade 8

Category	Measure	Ingredient	Cost/Student
Personnel			
	Forward	All Personnel Costs	$7.878
	MAP	All Personnel Costs	$5.850
	MCAP	All Personnel Costs	$4.188
Facilities		N/A	
Materials	Forward	All Materials Costs	$0.136
	MAP	All Materials Costs	$13.553
	MCAP	All Materials Costs	$6.890
Other		School Psychologist (data analysis)	$0.037
		SPSS Annual License	$0.813
Grand Total			$39.345

Table A.11

Personnel Ingredient Cost Breakdown

Method	Ingredient	Price (hourly)	Fringe Benefit Rate %	Total Price per Minute
Forward	Classroom Teacher	$44.77	52.02	$1.134
	Instructional Coordinator (DAC)	$31.87	54.08	$0.818
	Principle (SAC)	$48.80	54.08	$1.253
	District IT Manager (DTC)	$47.51	54.08	$1.220
	School IT Specialist (STC)	$30.51	52.02	$0.773
MAP	Classroom Teacher	$44.77	52.02	$1.134
	Instructional Coordinator (DAC)	$31.87	54.08	$0.818
	Principle (SAC)	$48.80	54.08	$1.253
	School IT Specialist (Systems Administrator)	$30.51	52.02	$0.773
	Administrative Assistant (Data Administrator)	$19.16	47.96	$0.472
MCAP	Classroom Teacher	$44.77	52.02	$1.134

Note. District Assessment Coordinator (DAC), School Assessment Coordinator (SAC), District Technology Coordinator (DTC), School Technology Coordinator (STC)

Table A.12

Forward Exam Personnel Ingredient Price x Time / Number of Students

Grade	Ingredient	Total Price per Minute	Real Minutes Required	Total	Number of Students	Cost per Student
6 & 7	Classroom Teacher (training)	$1.134	105	$34.030	26	$1.309
	Classroom Teacher (administration)	$1.134	30	$119.104	26	$4.581
	DAC (training)	$0.818	45	$36.829	7600	$0.005
	DAC (preparation and administration)	$0.818	2400	$1964.212	7600	$0.258
	SAC (training)	$1.253	45	$56.393	1587	$0.036
	SAC (preparation and administration)	$1.253	900	$1127.866	1587	$0.711
	DTC (training)	$1.220	90	$109.798	7600	$0.014
	DTC (preparation and administration)	$1.220	1200	$1463.976	7600	$0.193
	STC (training)	$0.773	90	$69.572	1587	$0.044
	STC (preparation and administration)	$0.773	600	$463.813	1587	$0.292
8	Classroom Teacher (training)	$1.134	115	$34.030	26	$1.309
	Classroom Teacher (administration)	$1.134	30	$130.447	26	$5.017
	DAC (training)	$0.818	45	$36.829	7600	$0.005
	DAC (preparation and administration)	$0.818	2400	$1964.212	7600	$0.258
	SAC (training)	$1.253	45	$56.393	1587	$0.036
	SAC (preparation and administration)	$1.253	900	$1127.866	1587	$0.711
	DTC (training)	$1.220	90	$109.798	7600	$0.014
	DTC (preparation and administration)	$1.220	1200	$1463.976	7600	$0.193
	STC (training)	$0.773	90	$69.572	1587	$0.044
	STC (preparation and administration)	$0.773	600	$463.813	1587	$0.292

Note. District Assessment Coordinator (DAC), School Assessment Coordinator (SAC), District Technology Coordinator (DTC), School Technology Coordinator (STC)

Table A.13
MAP Personnel Ingredient Price x Time / Number of Students

Ingredient	Total Price per Minute	Real Minutes Required	Total	Number of Students	Cost per Student
Classroom Teacher (training)	$1.134	60	$68.059	26	$2.618
Classroom Teacher (administration)	$1.134	45	$51.045	26	$1.963
DAC (preparation and administration)	$0.818	2400	$1964.212	7600	$0.258
SAC (preparation and administration)	$1.253	900	$1127.866	1587	$0.711
Systems Admin. (preparation and administration)	$0.773	600	$463.813	7600	$0.061
Data Admin. (preparation)	$0.472	3840	$1814.250	7600	$0.239

Note. District Assessment Coordinator (DAC), School Assessment Coordinator (SAC), District Technology Coordinator (DTC), School Technology Coordinator (STC)

Table A.14

MCAP Exam Personnel Ingredient Price x Time / Number of Students

Grade	Ingredient	Total Price per Minute	Real Minutes Required	Total	Number of Students	Cost per Student
6	Classroom Teacher (training)	$1.134	60	$68.059	26	$2.618
	Classroom Teacher (administration)	$1.134	8	$9.075	26	$0.349
	Classroom Teacher (scoring)	$1.134	26	$29.492	26	$1.134
7 & 8	Classroom Teacher (training)	$1.134	60	$68.059	26	$2.618
	Classroom Teacher (administration)	$1.134	10	$11.343	26	$0.436
	Classroom Teacher (scoring)	$1.134	26	$29.492	26	$1.134

Note. District Assessment Coordinator (DAC), School Assessment Coordinator (SAC), District Technology Coordinator (DTC), School Technology Coordinator (STC)

Table A.15

Diagnostic Accuracy of the Current Screening Practices and Minimally Intensive Changes, Adapted from Klingbeil et al. (2019)

Intensity of the Changes	Measure (Cut Score)	Grade	TP	FP	TN	FN	SE [95% CI]	SP [95% CI]
N/A (Current Practice)	MAP (Vendor)	6	117	24	321	49	.70 [.66, .74]	.83 [.79, .87]
		7	134	31	340	56	.71 [.67, .75]	.90 [.87, .93]
		8	187	24	254	50	.79 [.75, .83]	.84 [.80, .88]
Negligible (Data Available)	2016 Forward (Vendor)	6	107	27	317	60	.64 [.60, .68]	.92 [.90, .94]
		7	154	20	351	65	.66 [.62, .70]	.95 [.93, .97]
		8	186	20	258	51	.79 [.75, .83]	.93 [.91, .95]
Minimal	MAP (Local)	6	151	99	246	16	.90 [.87, .93]	.71 [.67, .75]
		7	173	69	302	17	.91 [.89, .93]	.81 [.78, .84]
		8	217	50	228	20	.92 [.90, .94]	.82 [.79, .85]
	2016 Forward (Local)	6	153	82	263	13	.92 [.90, .94]	.76 [.72, .80]
		7	171	90	281	18	.90 [.88, .92]	.76 [.72, .80]
		8	215	44	234	22	.91 [.89, .93]	.84 [.80, .98]
	2016 Forward & MAP	6	151	69	276	15	.91 [.89, .93]	.80 [.77, .83]
		7	173	70	301	17	.91 [.89, .93]	.81 [.78, .84]
		8	214	34	244	23	.90 [.87, .93]	.88 [.85, .91]

Note. n = 511, 561, 515 for grades 6, 7, and 8. Base rates of students scoring below proficiency equaled .32, .34, and .46 for grades 6, 7, and 8. Contingency table values based on pooled estimates from 20 imputation values (rounded to whole numbers). Confidence intervals for sensitivity and specificity were calculated using the formula presented by Harper and Reeves (1999). TP = True Positive; FP = False Positive; TN = True Negative, FN = False Negative; SE = Sensitivity; SP = Specificity

Table A.15 Continued

Diagnostic Accuracy of the More Intensive Changes, Adapted from Klingbeil et al. (2019)

Intensity of the Changes	Measure	Grade	TP	FP	TN	FN	SE [95% CI]	SP [95% CI]
More Intensive	MCOMP (Vendor)	6	66	5	340	100	.40 [.36, .38]	.99 [.98, 1.0]
		7	42	16	354	147	.22 [.19, .25]	.96 [.94, .98]
		8	103	10	268	134	.43 [.39, .47]	.96 [.94, .98]
	MCAP (Vendor)	6	57	11	334	109	.34 [.30, .38]	.97 [.96, .98]
		7	68	20	350	121	.36 [.32, .40]	.95 [.93, .97]
		8	104	11	268	133	.44 [.40, .48]	.96 [.94, .98]
	MCOMP (Local)	6	150	135	210	16	.90 [.87, .93]	.61 [.57, .65]
		7	173	172	199	16	.92 [.90, .94]	.54 [.50, .58]
		8	213	128	150	24	.90 [.87, .93]	.54 [.50, .58]
	MCAP (Local)	6	150	146	198	16	.90 [.87, .93]	.58 [.54, .62]
		7	171	183	188	19	.90 [.88, .92]	.51 [.47, .55]
		8	211	103	175	25	.89 [.86, .92]	.63 [.59, .67]
	MCOMP & MCAP	6	153	132	213	14	.92 [.90, .94]	.62 [.58, .66]
		7	171	166	204	19	.90 [.88, .92]	.55 [.51, .59]
		8	212	101	177	25	.90 [.87, .93]	.64 [.60, .68]
Most Intensive	2016 Forward & MAP & MCOMP	6	151	67	277	16	.91 [.89, .93]	.80 [.77, .83]
		7	172	71	300	17	.91 [.89, .93]	.81 [.78, .84]
	2016 Forward & MAP & MCAP	8	213	35	243	24	.90 [.87, .93]	.87 [.84, .90]

Note. n = 511, 561, 515 for grades 6, 7, and 8. Base rates of students scoring below proficiency equaled .32, .34, and .46 for grades 6, 7, and 8. Contingency table values based on pooled estimates from 20 imputation values (rounded to whole numbers). Confidence intervals for sensitivity and specificity were calculated using the formula presented by Harper and Reeves (1999). TP = True Positive; FP = False Positive; TN = True Negative, FN = False Negative; SE = Sensitivity; SP = Specificity

Figure A.1

Indices of Diagnostic Accuracy

		Predicted Performance on the Criterion Measure		
		Not Proficient	Proficient	
Actual Performance on the Criterion Measure	Not Proficient	True Positive (TP)	False Negative (FN)	Sensitivity $\frac{TP}{(TP + FN)}$
	Proficient	False Positive (FP)	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	NPV $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{TP + TN + FP + FN}$

Table A.16

Descriptive Statistics and Correlations for Each Measure, Adapted from Klingbeil et al. (2019)

	Grade		
	6	7	8
2017 Forward Exam (2017 FWD)			
n	497	550	506
M	645.33	666.72	667.54
SD	52.51	54.46	57.09
Skew	-0.63	-0.36	-0.64
Kurtosis	1.45	2.45	1.83
2016 Forward Exam (2016 FWD)			
n	454	528	454
M	633.69	655.17	655.10
SD	43.98	44.51	54.02
Skew	-.55	-0.001	-0.66
Kurtosis	1.64	1.18	-2.41
Measures of Academic Progress (MAP)			
n	509	558	515
M	230.83	239.07	241.89
SD	14.91	15.120	15.72
Skew	-0.38	-0.05	-0.21
Kurtosis	0.37	.735	0.26
Math Concepts and Applications (MCAP)			
n	489	541	492
M	21.61	17.95	14.02
SD	10.04	8.87	8.57
Skew	0.32	0.74	1.06
Kurtosis	-0.51	0.45	0.85
Correlations			
n	511	560	515
2017 FWD & 2016 FWD	.828	.816	.826
2017 FWD & MAP	.843	.831	.856
2017 FWD & MCAP	.701	.681	.707
2017 FWD & MCOMP	.714	.637	.718
MAP & MCAP	.757	.743	.770
MAP & MCOMP	.766	.683	.804
MCAP & MCOMP	.785	.720	.794

Note. 2016 Forward Exam scores represent students' performance in the previous grade. Correlations are based on pooled estimates from 20 imputation files. All correlations were statistically significant ($p < .001$).

CURRICULUM VITAE

SAMUEL A. MAURICE

EDUCATION

2021 (Anticipated)	**Ph.D.** Educational Psychology – School Psychology Specialization University of Wisconsin – Milwaukee (APA, NASP Accredited) Dissertation:
2017	**M.S.** Educational Psychology – School Psychology Specialization University of Wisconsin – Milwaukee (APA, NASP Accredited) Thesis:
2015	**B.S.** Psychology Bemidji State University

PRE-DOCTORAL INTERNSHIP

2020-2021	**Pre-Doctoral Internship in Psychology** **Boys Town Center for Behavioral Health** Nebraska Internship Consortium in Professional Psychology *Pre-doctoral APPIC internship accredited by the American Psychological Association*

PSYCHOLOGY PRACTICUM

2019-2020	**Milwaukee College Preparatory School** Advanced School Psychology Focused Practicum
2018-2019	**Roger's Behavioral Health** Advanced Clinical Focused Practicum
2017-2018	**Family Options Counseling** Clinical Focused Practicum
206-2017	**Greenfield School District Department of School Psychology** School Psychology Focused Practicum

TEACHING EXPERIENCE

2019-2020 **Associate Lecturer**
Introduction to Learning and Development (Ed Psy 330)
University of Wisconsin – Milwaukee

2014-2015 **Teaching Assistant**
Lifespan Development (Psy 3237)
Bemidji State University

RESEARCH EXPERIENCE

2016-2020 **Research Assistant**
Research team of Dr. David Klingbeil, Ph.D.
University of Wisconsin – Milwaukee

PUBLICATIONS

Cipriano, D., **Maurice, S. A.**, & Chayer, R. (under review). School-Based Mental Health Impacts Academic Outcomes: A Two-Year Longitudinal Study on an Urban Population. Manuscript under review for publication in School Mental Health.

Klingbeil, D. A., **Maurice, S. A.**, Van Norman, E. R., Nelson, P. M., Birr, C., Hanrahan, A. R., Schramm, A. L., Copek, R. A., Carse, S. A., Koppel, R. A., & Lopez, A. L. (2019). Improving Mathematics Screening in Middle School. *School Psychology Review*, *48*(4), 383–398. https://doi.org/10.17105/SPR-2018-0084.V48-4

Kwon, K., Teer, J. E., **Maurice, S. A.**, & Matejka, C. M. (2020). Self-report of emotional experience and peer nominations of expressivity: Predictability of change in teacher-rated social behavior. *Social Development*, *29*(3), 837–853. https://doi.org/10.1111/sode.12429

Klingbeil, D.A., **Maurice, S.A.**, & Schramm, A.L. (2019). Legal and ethical considerations for providing behavior interventions in schools. In K.C. Radley & E.H. Dart (Eds.). Handbook of Behavioral Interventions in Schools: Multi-Tiered System of Support. New York, NY: Oxford University Press.

PRESENTATIONS

Cipriano, D.J., **Maurice, S.A.**, Bauernfeind, C., Watkins, J. (2019, May). School-Based Mental Health Care Impacts Academics for Children in Marginalized Communities. Poster presented at the annual Engage conference of the Medical College of Wisconsin. Milwaukee, WI.

Maurice, S.A., Schramm, A.L., & Karlak, J.J. (2019, February). Evaluating the Cost-Effectiveness of Two Social-Emotional Interventions. Poster presented at the annual convention of the National Association for School Psychologists. Atlanta, GA.

PRESENTATIONS (CONT.)

Maurice, S.A., Schramm, A.L., & Hanrahan, A.R. (2018, February). Balancing accuracy and efficiency: Incremental validity of math screening measures. Paper presented at the annual convention of the National Association for School Psychologists. Chicago, IL.

Klingbeil, D.A., Van Norman, E.R., **Maurice, S.A.**, Schramm, A.L., & Birr, C (2017, August). Accuracy of universal screening cut-scores across changes in state assessments. Poster presented at the annual convention of the American Psychological Association. Washington, DC.

Maurice, S.A. & Karlak, J.J. (2017, February). Using social validity to culturally modify evidence-based interventions. Poster presented at the annual convention of the National Association for School Psychologists. San Antonio, TX.

Larson, K. & **Maurice, S. A.** (2014, October). Cell phone use in the classroom: Factors, frustration, and facilitating change. Presentation at the annual convention of the International Society for Exploring Teaching and Learning. Denver, CO.

WORK EXPERIENCE

2016-2020 **Research Assistant**
Consulting Office for Research and Evaluation
University of Wisconsin – Milwaukee

2015-2016 **Graduate Assistant**
Next Door Foundation
University of Wisconsin – Milwaukee

SERVICE

Student Reviewer
Assessment for Effective Intervention

Student Reviewer
National Association of School Psychologists

Professional Association Membership
National Association of School Psychologists