Fall 2018

Evaluating the Criterion Validity and Classification Accuracy of Universal Screening Measures in Reading

Asia Thomas

EVALUATING THE CRITERION VALIDITY AND CLASSIFICATION ACCURACY OF UNIVERSAL SCREENING MEASURES IN READING

by

Asia Thomas

Bachelor of Arts
University of Georgia, 2015

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Arts in

School Psychology

College of Arts and Sciences

2018

Accepted by:

Stacy-Ann A. January, Director of Thesis

Nicole Zarrett, Reader

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

ABSTRACT

Educators use universal screening in the context of Response to Intervention frameworks to identify students who may be at risk for not meeting proficiency on the state assessment. Given the potential high-stakes of state tests, using accurate screening systems is critical for early remediation. Independent research is emerging on comprehensive and expensive reading screeners such as the Measures of Academic Progress (MAP), a computer adaptive test, and the Strategic Teaching Evaluation of Progress (STEP), a developmental reading assessment. The current study evaluated the criterion-related validity of MAP and STEP with a state assessment. Additionally, the utility of each screening measure to distinguish between students at risk and not at risk for reading failure was evaluated. Participants were two cohorts (Cohort 1 *N* = 209; Cohort 2 *N* = 115) of children enrolled in a public charter school system. MAP and STEP were administered in spring of second grade, and fall and spring of third grade. Results suggested that MAP and STEP scores were significant strong predictors of third-grade state assessment scores. Hierarchical regression analyses indicated that STEP scores explained more variance in scores on the state assessment above and beyond MAP scores alone; however, this effect was generally small. Furthermore, findings support the utility of MAP and STEP in distinguishing between students at risk for not meeting reading proficiency. Altogether, results from this study support the use of both of MAP and STEP scores to predict reading performance. However, MAP alone may be sufficient as a single screening measure.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Negative and potentially long-term consequences across the lifespan are possible for students who have early reading difficulties (Darney, Reinke, Herman, Stormont, & Ialongo, 2013; January et al., 2017). Struggling readers who do not receive early remediation are at increased risk for dropping out of high school, suicide ideation, future homelessness, teenage pregnancy, and juvenile delinquency (Bennett, Frasso, Bellamy, Wortham, & Gross, 2013; Daniel et al., 2006; January et al., 2017; McGill-Franzen, 1987; Vitaro, Brendgen, Larose, & Tremblay, 2005). Educational legislation such as the Every Student Succeeds Act (ESSA) and Individuals with Disabilities Education Improvement Act (IDEA), were implemented to address national concerns regarding reading failure and to prevent poor academic outcomes (ESSA, 2015; IDEA, 2004). Many schools have adopted Response to Intervention (RtI) frameworks using early student performance data to inform multi-tiered systems of instructional support. Within the context of RtI, students who display early signs of risk for reading difficulties receive targeted interventions to promote reading proficiency (Kettler, Glover, Albers, & Feeney-Kettler, 2014).

In school districts across the nation, educators conduct universal screening in reading as an essential component of an RtI framework to inform appropriate instructional placement and to prevent poor academic trajectories (Jimerson, Burns, & VanDerHeyden, 2016). Universal screening measures are brief assessments used to identify students with some risk in the area of reading (Catts, Nielsen, Bridges, Liu & Bontempo, 2015; Glover & Albers, 2007). Screening typically consists of three periods (i.e., fall, winter, spring) of assessing skills that are potential intervention targets and highly predictive of future reading outcomes (Jenkins, Hudson, & Johnson, 2007). Measures of reading proficiency should also align with the National Reading Panel's five components of reading instruction: phonological awareness, alphabetics, vocabulary, fluency, and reading comprehension (U.S. Department of Health and Human Services, 2000). Given the urgency of accurately identifying students who may be at risk for reading difficulties (Catts et al., 2015; Vaughn et al., 2008), practitioners and researchers need universal screeners that are valid measures of reading proficiency with adequate technical utility.

Researchers describe validity as an integrated evaluation of empirical evidence and theoretical rationales, to support the competence and utility of a measure for data-based decision making (Messick, 1989). Validity is a comprehensive quality logically comprised of many different sources of substantiation (Nelson, 2009). Some of the critical features for educators to consider when evaluating the utility of a screener include criterion validity, classification accuracy, efficiency, and consequential validity (Jenkins et al., 2007; Jenkins & Johnson, 2009; Messick, 1989). The present study examines the criterion validity and classification accuracy of two widely used universal screening measures.

Critical Features of a Screener

Criterion validity refers to the extent to which a screener correlates with an established criterion measure of reading that is either norm-referenced or criterion-referenced (Glover & Albers, 2007; Jenkins et al., 2007). A norm-referenced criterion measure is used to assess where students rank among their peers within an area of achievement and helps to distinguish between low and high performers. Criterion-referenced tests are used to determine whether students achieve specific learning goals and can be informative when evaluating student performance before and after instruction has finished (Aiken, 1985).

Measures that assess the same or similar constructs should have a strong association (Jenkins et al., 2007). The association between measures of the same construct demonstrates the psychometric characteristics of either concurrent validity, when measures are completed at the same time point, or predictive validity, when there is a lag between assessment administration (Glover & Albers, 2007; Jenkins et al., 2007). Although there are no recognized criteria to interpret correlations, coefficients of .40 and above indicate an existing relation between two measures, assuming the criterion measure is valid (Burns, Haegele, & Petersen-Brown, 2014). Criterion validity is an essential first step to support technical utility; however, correlations between a screener and criterion measure only represent the potential of a screener to accurately distinguish between students with either inadequate or proficient performance (Jenkins et al., 2007), examining classification accuracy is also critical.

Classification accuracy refers to the extent to which a universal screener accurately identifies students as “at-risk,” based on their criterion measure performance (Jenkins et al., 2007). Students who score below the grade-level benchmark on a screening measure are considered “at-risk” for poor performance on a criterion measure, whereas students who score above benchmark are “not at-risk.” When creating benchmarks, test developers typically reverse their steps—first assessing student performance on a criterion measure, then pinpointing the score on the screening measure that distinguishes among students who passed the criterion measure and those who did not (Jenkins & Johnson, 2009). The pinpointed score becomes the benchmark students are expected to reach at a specific screening period. This benchmark norm is useful for predicting student performance on subsequent criterion measures. Accurate prediction of student performance on a criterion measure is a first step in helping school personnel make appropriate instructional placement decisions (Ball & Christ, 2012; Kettler et al., 2014).

Classification accuracy can be described as the conditional probability of student outcomes characterized by true or false positives and negatives. Once test developers establish benchmarks, screening assessments indicate a dichotomous outcome with four probable scenarios: true positives, true negatives, false positives, or false negatives (Riedel, 2007). False positives occur when screeners identify a student as being at risk for performing poorly on an outcome measure, when in reality they were not at risk and did not fail the outcome measure. False positives are undesirable, as they can result in the misallocation of time and resources (Jenkins & Johnson, 2009). False negatives occur when screening results indicate that there was no risk, despite actual risk and subsequent failure on the outcome measure. False negatives are also problematic because students who are truly at risk may not receive appropriate support, and may fall further behind and experience long-term consequences. Additionally, true positives occur when screeners accurately identify students at risk, and true negatives occur when screeners accurately identify students who are not at risk. Screeners that err on the side of making more false positives as opposed to false negatives have the most utility (Christ & Nelson, 2014). Avoiding false negatives and yielding a high percentage of true positives is an imperative component of data-based decision making.

Additionally, when evaluating the classification accuracy of a screening measure, a sensible balance between sensitivity or the proportion of students correctly identified as having some risk in reading, and specificity, those correctly identified to have proficient reading skills, is desirable. Sensitivity [i.e., true (+)/true (+) + false(-)], is deemed acceptable at levels greater than 90 and specificity [i.e., true (-)/true (-) + false (+)], should be at least 80% (Catts et al., 2015; Jenkins & Johnson, 2009). Although not all researchers agree about optimal levels of sensitivity and specificity set at 90% and 80% respectively, there is a trade-off for decreased levels of specificity with any increase in level of sensitivity, and vice versa (Glover & Albers, 2007; Jenkins & Johnson, 2009).

Test-developer suggested benchmarks may or may not be tailored to fit the needs of a specific school or population of students. As an alternative, locally derived cut-scores from Receiver Operating Characteristic (ROC) curve analyses may be used to obtain accurate predictions of student performance (Klingbeil, Nelson, Van Norman, & Birr, 2017; Walker, Small, Severson, Seeley, & Feil, 2014). The area under the curve (AUC) value obtained from ROC curve analyses represents the overall accuracy of a screening measure ranging in value from no predictive value, 0.0, to perfect accuracy, 1.0 (Christ & Nelson, 2014; Walker et al. 2014). While maintaining a balance between levels of sensitivity and specificity and reaching acceptable AUC values, it is important that benchmarks are contextually appropriate regardless of the methods used to create them (Klingbeil et al., 2017; Nelson, Van Norman, & Lackner, 2017).

Universal Screening Measures of Reading

One widely used screening assessment, curriculum-based measurement of oral reading (CBM-R), is a brief measure of reading rate and accuracy with more than three decades of research supporting its psychometric properties (Deno, 1985; January & Ardoin, 2015; January, Ardoin, Christ, Eckert, & White, 2016). In support of its utility, the relation between CBM-R and standardized tests of reading performance is characterized by an extensive literature-base as moderately high (Reschly, Busch, Betts, Deno, & Long, 2009; Wayman, Wallace, Wiley, Tichá, & Espin, 2007). An accumulation of research also supports the diagnostic qualities of CBM-R as a valid predictor of student risk for reading difficulties (Kilgus, Methe, Maggin, & Tomasula, 2014; Yeo, 2010). Less research has been conducted, however, to validate other frequently used and more comprehensive screening tools such as Strategic Teaching and Evaluation Progress, (STEP; Urban Education Institute [UEI], 2011) a developmental reading assessment and the computer adaptive measure, Measures of Academic Progress (MAP; Northwest Evaluation Association [NWEA], 2011). To date, no published research exists on STEP, despite its widespread use in about 18 states (UEI, 2011). Additionally, peer-reviewed, published evidence to support the technical adequacy of MAP is emerging (e.g., Ball & O'Connor, 2016; January & Ardoin, 2015; Klingbeil et al., 2015; Klingbeil et al., 2017). In comparison to CBM-R, MAP and STEP are more expensive and take longer to administer. However, given the broad range of literacy skills assessed, MAP and STEP may better inform educators about a student's pattern of strengths and weaknesses across essential components of reading instruction. It is therefore of interest to researchers and school personnel to evaluate the validity and utility of these alternative screening measures administered to a large population of students across the nation.

Measures of Academic Progress. The Measures of Academic Progress (MAP; Northwest Evaluation Association [NWEA], 2011) is a group-administered computer adaptive test of academic performance used to measure growth, predict student performance, and inform instruction. The MAP reading assessment covers a broad range of foundational skills that are necessary for successful reading at different grade levels. In second and third grade, the reading composite measures phonological awareness and vocabulary, as well as literal, interpretive, and evaluative comprehension. Despite that MAP is a screener that directly measures a range of reading skills, it is expensive and takes students 30 to 90 minutes to complete. However, MAP is widely-used, as it is administered to over ten million students across the nation.

Guided by Item Response Theory, MAP factors in the difficulty of each test item, as well as each student's unique ability when presenting individual assessment questions (NWEA, 2015). First, the probability of a student answering correctly on a particular item is estimated. Actual student performance is then used to adjust the difficulty of all successive items, given the responses to previous test questions. This valuable characteristic of test adaptability provides both a challenge for the highest performing students and also prevents lower performing students from becoming frustrated by questions that are too difficult (NWEA, 2015). Tailoring test questions to individual levels of achievement as opposed to a static bank of questions that could be above or below a student's skill level makes for a more accurate measure of student performance (NWEA, 2015). Scores on the MAP are conveyed as a Rasch Unit (RIT) score, or stable, equal interval scales that allow researchers and practitioners to compare student growth across grades (NWEA, 2015). The RIT score that a student receives upon test completion represents the level of test item difficulty they can answer correctly about 50% of the time (NWEA, 2015). Reports from MAP's technical manual include marginal reliability ranges of .94 to .95 for Grades 2 through 5, test-retest reliability ranges of .70 to .85 for students in Grades 2 to 10, internal consistency of .70 to .86 across grades, and concurrent validity estimates of .57 to .82 across states (NWEA, 2011).

The NWEA (2016) recently conducted a study using a large sample of 1,615 students in third grade across 23 schools in Georgia (GA). Concurrent MAP administration in the spring indicated a strong association (r = .83) with Georgia Milestones (a computerized, high-stakes state assessment), supporting the validity MAP scores. Although NWEA provides information regarding the technical adequacy of MAP (e.g., NWEA, 2016), additional research is needed to support the technical utility of MAP for the purpose of identifying students at risk. Independent peer-reviewed evidence is emerging in this area, and a few recently-published studies have investigated MAP's technical characteristics.

Ball and O'Connor (2016) evaluated the predictive validity and classification accuracy of MAP scores in the spring of second grade with a state assessment in the fall of third grade. They reported strong correlations (r = .82) among the measures, and created locally derived cut-scores. In one sample, sensitivity was held at .95 and specificity was .58, with 62% of correct classifications reported. For the second sample, sensitivity was 1.00 and specificity was .53, with 59% of correct classifications indicated. However, results of this study are limited to the specific criterion measure used (i.e., state test in Wisconsin) and the homogenous sample of students with respect to race/ethnicity and socioeconomic status. Additionally, the locally derived cut-scores may not generalize to lower performing students given the low base rates (11%) of failing the state test (Ball & O'Connor, 2016). Research is needed with samples of differing demographic characteristics, such as those students of racial/ethnic minority background and who live in low-resource communities.

In a study with second- and third-grade students, Klingbeil, McComas, Burns, and Helman (2015) examined the criterion-related validity and classification accuracy of fall MAP scores in predicting end-of-year reading proficiency using spring MAP scores. Strong correlations (r = .77) between fall and spring MAP were reported, in addition to sensitivity and specificity levels of .85 and .81, respectively, using test-developer benchmarks. In a recent study using publisher-recommended cut scores, Klingbeil, Nelson, Van Norman, and Birr (2017) found strong correlations (r = .74) between fall MAP scores and a spring-administered state assessment for students in Grade 3. Sensitivity and specificity levels were .63 and .83, respectively, in third grade. Finally, January and Ardoin (2015) found a high correlation (r = .87) of MAP scores with the concurrently administered norm-referenced Iowa Test of Basic Skills (ITBS) in third grade.

In sum, previous research suggests that statistically significant and strong associations exist between MAP and other assessments of reading achievement. MAP also has growing evidence to support its classification accuracy. However, replication of these findings is needed with students that are different in terms of race and ethnicity, region, type of school (i.e., public, charter), base rates, and socioeconomic background.

Strategic Teaching and Evaluation of Progress. Over 230 schools in 29 cities currently use the Strategic Teaching and Evaluation of Progress, (STEP; Urban Education Institute [UEI], 2011) developmental reading assessment, and test developers predict an increase in numbers. The STEP assessment tool is a formative evaluation of reading administered to students in pre-K through third grade at four points across the school year. Assessment results from STEP provide a detailed profile of students' strengths and weaknesses in a specific set of skills used for informing instruction. Furthermore, the STEP tool includes a set of age-appropriate texts that increase in difficulty with each "step," used to assess students' reading rate, accuracy, prosody, and reading comprehension. STEP also pairs the reading of each leveled book with an assessment of letter-sound knowledge, phonological awareness, concepts about print, and spelling during each administration (Kerbow & Bryk, 2005).

Like MAP, STEP is expensive, and there are schools that administer both MAP and STEP for universal screening. STEP can also be resource intensive, as it is administered during individual student conferences. Then the administrators must input, analyze, and use student data to inform both instruction and student risk status. Despite the time and costs associated with administration, educators like STEP because results include a detailed profile of students' strengths and weaknesses on a specific set of skills that can be used for tailoring instruction. During each one-to-one student evaluation, administrators gain a personalized understanding of a student's reading ability. This individualized approach helps to inform ongoing guided reading instruction and has the potential to translate directly to delivery of timely intervention. Despite the added value that STEP brings to ongoing reading instruction and timely intervention, there is no known independent research to support its utility and accuracy as a screening measure. Additionally, there is no peer-reviewed evidence to support the added value of STEP to screening measures that are already supported by research like MAP. STEP, however, continues to be used by schools across the nation without continued independent empirical validation to support its utility and adequacy.

In addition to its potential utility for guiding instruction, test developers posit that the STEP tool can be used to predict student performance on high-stakes assessments (Kerbow & Bryk, 2005). In many states, third grade is the first time that students take high-stakes standardized tests that, among other purposes, may be used to determine promotion to fourth grade (Weyer, 2018). Students in third grade are expected to move from STEP level 10 to 12. According to the STEP technical report, students who successfully achieve the benchmark of STEP 12 by the end of third grade are more likely to perform at or above proficiency on standardized tests of reading (Kerbow & Bryk, 2005). By STEP 12, publishers anticipate that students can acquire new meaning of words and concepts through self-sufficient reading, understand texts dissimilar from their own relatable experiences, and develop the ability to consider multiple story interpretations. The higher-order skills anticipated at STEP 12 are all in addition to previously mastered foundational skills at earlier STEP levels (Kerbow & Bryk, 2005).

The STEP technical report provides evidence supporting both reliability and validity. The sample used to validate STEP as a screening measure included four cohorts of African American students from a single Chicago public school, and approximately 75% of students were of low-income background. To measure concurrent validity, Kerbow and Bryk (2005) compared students' spring STEP performance with the ITBS and the Illinois Standards Achievement Test (ISAT) in third grade. Results of the study revealed high correlations between STEP in relation to ITBS (*r* = .60) and ISAT (*r* = .66). When measured concurrently with the Degrees of Reading Power, there was a .51 correlation with STEP in the spring of first grade and a .62 correlation in the spring of second grade. STEP scores in the spring of second grade significantly predicted STEP scores in third grade on the ITBS (*r* = .58) and ISAT (*r* = .68; Kerbow & Bryk, 2005). Based on these cohorts of children, STEP appears to demonstrate moderate to strong psychometric properties, but independent research is necessary to supports its technical characteristics.

In addition to criterion-related validity, test developers reported a negative predictive value of 80% for both second- and third-grade students (Kerbow & Bryk, 2005). The NPV of a measure represents the percentage of students who are identified by a screener as not being at risk, who were observed to perform proficiently on the criterion measure as predicted. In contrast, positive predictive value is the percentage of students identified by a screener for being at risk, who did not meet proficiency standards as projected (Kilgus et al., 2014). The NPV represents the percentage of students in second and third grade who met their respective grade level benchmarks on the STEP assessment and later scored at or above the 50th percentile on the ITBS. Furthermore, 86% of third-grade students who met the STEP level benchmark in the spring reached or surpassed state standards on the ISAT (Kerbow & Bryk, 2005). In sum, a majority of students who met the targeted STEP benchmarks passed their standardized state assessment. Despite these findings, the extent to which these specific benchmarks accurately predict performance on different outcome measures other than those included in the STEP technical report is unknown. An investigation of the consistency of these results across different criterion measures of reading, as well as with populations in regions other than the Midwest is warranted. Additionally, the utility of these benchmarks for populations other than those included in the technical report are unknown.

Summary

Identification of screening measures that are psychometrically sound and backed by empirical evidence are necessary to provide educators and researchers with recommendations for identifying students who may be at risk. Although shorter and less expensive screeners like curriculum-based measurement of oral reading are available and have strong psychometric properties (e.g., January et al., 2016), many schools choose to use measures such as MAP and STEP. Nonetheless, longer screeners may be of value to educators who desire comprehensive assessments that, in addition to identifying students who may have some risk in reading, may be useful for informing instruction by providing a detailed snapshot of the skills student have and have not mastered.

Despite its use in schools, no published research exists on the STEP tool, and empirical research on the technical adequacy of MAP is growing (e.g., Ball & O'Connor, 2016; January & Ardoin, 2015; Klingbeil et al., 2015, 2017). Furthermore, schools may administer both a developmental reading assessment, like STEP, and a computer adaptive test, like MAP, for the purposes of universal screening. The potential benefit of administering both measures for universal screening is unknown. Thus, it is necessary to determine whether administering both MAP and STEP for universal screening is worth the allocation of time and resources. This information will help educators decide which measures are useful for the purposes of universal screening, and which might be better suited for informing instruction. It is also critical that educators have accurate screening measures to make informed decisions for students in need of immediate intervention before state testing since performance becomes a factor that helps determine grade promotion. Additionally, there is an ongoing need to investigate the technical adequacy of screeners with samples of students who differ across demographic and school performance level variables from the extant peer-reviewed studies. Specific to STEP, more recent information on its technical characteristics is needed.

Current Study

The purpose of this study was to address current gaps in the literature regarding the technical utility of two common and comprehensive screening tools administered simultaneously in the same district. This study examined the concurrent and predictive validity as well as the classification accuracy of each assessment. Given that schools administer both measures at the same time, we were interested in whether adding STEP to MAP explained additional variance in students' reading achievement. For classification accuracy, statistically-optimized cut-scores were compared to the publisher-provided benchmarks for MAP. Publisher-recommended benchmarks for STEP were examined to evaluate the adequacy of each assessment to differentiate between students at risk for not meeting proficiency standards. Scores from two cohorts of second-grade students (who also had data from their third-grade year) were used, due to the increased focus on assessment outcomes for students taking high-stakes state tests for the first time in third grade. The following research questions were addressed in this study:

1. What is the concurrent and predictive validity of MAP and STEP with the high-stakes state assessment?

2. What is the incremental validity, if any, of administering STEP with MAP to predict performance on the high-stakes state assessment?
3. What is the classification accuracy of using publisher-recommended vs. statistically-optimized cut-scores when using MAP scores to predict state test performance?
4. What percentage of students who meet spring second- and third-grade benchmarks for STEP meet proficiency on the state assessment in third grade?

CHAPTER 2

METHOD

2.1 PARTICIPANTS AND SETTINGS

Extant data were obtained from two public charter schools located in an urban school district in the southeast region of the United States via a partnership between the school district and the researchers. All personally-identifiable student information was removed from the dataset by school personnel before it was provided to the researchers and replaced with unique ID numbers. Therefore, the university's Institutional Review Board did not consider this study to be human subjects research. The dataset included the assessment results for two consecutive cohorts of second-grade students (i.e., Cohort 1 and Cohort 2) who were also assessed during third grade. An analysis of variance (ANOVA) was conducted to determine whether the means of the two cohorts were statistically different. These analyses revealed that performance on the high-stakes assessment in Cohort 1 ($M = 504$) was statistically lower than performance of Cohort 2 ($M =522$), in addition to significant mean differences for most assessment scores. Therefore, cohorts were analyzed separately, and findings from Cohort 2 were used to investigate potential replication of findings from Cohort 1.

Across cohorts, data for 347 students were obtained. Based on school demographic data, the majority of participants were from low-income backgrounds; individual socioeconomic status data were not available. Demographic information for both cohorts were relatively similar across variables, and are summarized in Table 2.1. In

Cohort 1, somewhat greater proportion of males were included, whereas Cohort 2 had slightly more females (gender missing for 4.9% in Cohort 2). Majority of participants were Black/African American (97.8% in Cohort 1 and 92.6% in Cohort 2), or Hispanic/Latino (2.2% in Cohort 1 and 1.6% in Cohort 2). The number of students with an Individualized Education Program across cohorts was 23 (data missing for six students), and one student had limited English proficiency (data missing for seven students). In Cohort 1 and 2 respectively, 11% of the 1, 254 data points and 2% of the 690 data points of MAP and STEP scores were missing across grade and season for reasons unknown. A pairwise present approach to missing data was used, which led to varied sampled sizes for each analysis.

2.2 MEASURES

The MAP (NWEA, 2011) is a multiple-choice, group administered computer adaptive test used to measure student growth across three time points during the school year. The MAP assessment reading items cover a broad range of sub-skills required for effective reading at different grade levels, including phonological awareness, vocabulary, and comprehension. Key content areas of the assessment include (a) word analysis and vocabulary development, (b) literary response and analysis and, (c) reading comprehension (NWEA, 2011). The difficulty of each question is based on each individual student's performance on previously answered test items making the overall test adaptable to students' achievement level. Scores on the MAP are reported as RIT scores, an equal interval scale which estimates student achievement based on the difficulty of individual items (NWEA, 2015). Although tests are not timed and the number of questions each individual student sees may vary, students typically complete the MAP assessment in approximately 30 to 90 minutes.

The 2015 NWEA RIT scale study established recommended benchmark norms or RIT scores of 188.7 during spring administration in second grade and scores of 188.3 and 198.6 during third grade fall and spring administration correspondingly. Reports from MAP's technical manual include marginal reliability ranges of .94 to .95 for Grades 2 through 5, test-retest reliability ranges of .70 to .85 for students in Grades 2 to 10, and internal consistency of .70 to .86 across grades. (NWEA, 2011).

The STEP (UEI, 2011) assessment tool is typically administered to individual students four times each school year. Students may achieve more than one step depending on the growth they demonstrate during each assessment. Observation of reading behavior and prompting for understanding of the texts students read usually takes 10-15 minutes. In addition to observed reading, there is an underlying skills assessment with timed and untimed components.

STEP is designed to measure skills that contribute to overall mastery of reading as evidenced by scientifically recognized milestones (Kerbow & Bryk, 2005). The STEP tool is organized into a developmentally sequenced set of tasks used to inform instructors about the age-related patterns of individual students in comparison to same-grade peers. STEP interlaces components of phonemic awareness, concepts about print, reading fluency, and comprehension into one diagnostic assessment. Overall scale reliability was .98 (Kerbow & Bryk, 2005). As mentioned previously, concurrent and predictive validity with norm-referenced reading achievement tests generally fall within the moderate range (Kerbow & Bryk, 2005).

The Georgia Milestones is an end-of-grade summative, computerized test designed for Grades 3 through high school. Classified as a high-stakes assessment, it provides information to help teachers determine how ready students are to advance to the following grade. It is used to measure how well students have mastered state content standard skills in English Language Arts after which the state assessment was developed. The most recently revised standards for 2015-2016 in third grade include the following: (1) knowing and applying grade-level phonics and word analysis skills in decoding words, and (2) reading with sufficient accuracy and fluency to support comprehension. Students in third grade may have up to 240 minutes to finish three sections of the test which include two domains: Reading and Vocabulary in addition to Writing and Language. The assessment has a total of 60 items which include multiple choice, written response, and essay writing. There are also two kinds of essays, an opinion essay and an informational or explanatory essay (Georgia Department of Education, 2015).

Students' standard scores for the state assessment were used as the criterion-referenced measure for the current study with the following competence stage classifications: 'proficient' scores meeting standard (i.e., scale score range of 525-580), 'distinguished' scores exceeding standard (i.e., 581-830), 'beginning' scores (i.e., 180-474) and 'progressing' scores (475-524) both considered to be below standard. State law requires students in third grade to earn an at or above grade level competency stage classification to be promoted to fourth grade (Georgia Department of Education, 2015). The Georgia Department of Education reports a test reliability range of .85 to .94.

2.3 PROCEDURES

MAP and STEP were administered across three cycles in the fall, winter, and spring of each school year. MAP was administered as a group assessment using school computers that were available either in a computer lab or students' classrooms. The STEP assessment had a large testing window in the winter period that varied between schools. Thus, only fall and spring data were analyzed as these windows were concurrent for both assessments. Classroom teachers or school administrators collected STEP screening data during individual student conferences. The school districts' model of screening allowed for whole class supervision while a partner teacher or school administrator conducted the STEP assessment to avoid interruption of instructional time. Fidelity of implementation for the STEP assessment was not available; however, test administrators received on-site training from a STEP assessment representative who helped with initial planning and training sessions, as well as delivery of ongoing support for effective implementation. Finally, the state test was administered in the spring of third grade using school computers.

2.4 DATA ANALYSIS

All statistical analyses were conducted using SPSS v25. Preliminary descriptives, statistical assumptions, and test of normality were assessed for all variables. Generally, Pearson's correlations were conducted; however, Spearman's rho correlations were conducted for any analyses using STEP, due to the ordinal nature of the data. Next, hierarchical linear regression was used to determine the shared and unique variance of screening measures in relation to the Georgia Milestones. MAP was entered first because it shares the characteristic of being a computer-administered assessment, is convenient to employ, is supported by an emerging peer-reviewed literature base (e.g., Ball & O'Connor, 2016; January & Ardoin, 2015; Klingbeil et al., 2017), and has content overlap with the Georgia Milestones.

Receiver Operating Characteristic (ROC) curves and logistic regression were used to assess classification accuracy of MAP using the Georgia Milestones as the criterion measure. High stakes assessment scores below 525 were coded 1 for non-proficiency, and scores at or above 525 were coded 0 for proficiency. Statistically-optimized cut-scores for each administration of the MAP assessment were created to report the conditional probability of student outcomes. Due to the categorical nature of STEP data (i.e., the ordinal values of each STEP level), classification accuracy data were hand-calculated. First, all student STEP scores were coded into conditional probabilities of true or false positives and negatives. Students who scored at or above STEP level 9 in second grade or STEP level 12 in third grade and subsequently scored at or above 525 on the state test, were coded as 1 for True Negative. Students who scored below their grade level benchmark and did not meet state assessment proficiency were coded as 2 for True Positive. Students who scored at or above their grade level benchmark but scored below 525 were coded as 3 for False Negative. Students who scored below their grade level benchmark but scored at or above 525 were coded as 4 for False Positive. Each conditional probability was totaled in order to compute sensitivity, specificity, NPV, and PPV using programmed Microsoft excel formulas [i.e., sensitivity = true (+)/true (+) + false (-), specificity = true (-)/true (-) + false (+), NPV = true (-)/ true (-) + false (-), PPV = true (+)/ true (+) + false (+)]. Fall third grade STEP benchmarks were not provided by the test developers and therefore were not of interest during the classification accuracy portion of the study. Further, although all of these results will be presented, we will focus on NPV, as this is what was reported in the technical manual.

Table 2.1

Demographic Information for Participants

Cohort	*n*	Female	Male	LEP	SPED	Black or African American	Hispanic or Latino	More than one Race
1	225	115	110	1	15	220	5	–
2	122	53	63	–	8	113	2	1

Note. LEP = limited English proficiency; SPED = special education

CHAPTER 3

RESULTS

3.1 DESCRIPTIVE STATISTICS

Following preliminary analyses, all variables were normally distributed as assessed by Shapiro-Wilk's test (p > .05) with no significant outliers. All statistical assumptions tests resulted in nonsignificant violations since regression is robust to issues of non-normality. Table 3.1 provides a summary of descriptive statistics including the sample size, mean, and assessment correlations for each cohort of participants.

3.2 INFERENTIAL STATISTICS

Concurrent and predictive validity of MAP and STEP

Results from Cohort 1 indicated that scores for each screening measure had statistically significant (p < .01) associations with the GA Milestones. The highest predictive correlation was between second-grade spring MAP scores and third-grade state assessment scores (r = .82). The range of spring correlations for both grades were higher (r = .78 to .82), compared to the fall correlations in third grade (r = .75 to .77). STEP correlations across time points ranged from .77 to .78, whereas MAP correlations were between .75 and .88. In Cohort 2, all screening measures had statistically significant (p < .01) associations with the GA Milestones. Specifically, third-grade fall MAP scores resulted in the highest correlation r = .77. Overall, spring third grade scores demonstrated a slightly larger association with the state assessment in comparison to the spring second-grade scores. Across cohorts, the associations between screening measures and the outcome measure were moderate to large, and ranged from .62 to .82 across cohorts. Across all time points, MAP scores were a stronger predictor of the state assessment than STEP scores.

Incremental validity of administering STEP with MAP

Results of the hierarchical regression analyses are displayed in Table 3.2. In Cohort 1, MAP scores alone explained between 52 and 65% of the variance in third-grade state test performance across time periods. The addition of STEP scores led to statistically significant but small increases in the percent of variance explained (ΔR^2 range = .07 to .13) across models. The greatest increase (ΔR^2 =.13; $F(1, 95) = 36.019$, $p <$ .001) was when screeners were administered in fall of third grade. In Cohort 2, MAP scores explained 43 to 59% of the variability in state assessment scores. Similar to Cohort 1, STEP performance added an additional 4 to 6% of variance explained. Unlike Cohort 1, the largest ΔR^2 of .08, $F(1, 105) = 17.705$, $p < .001$, was observed when screeners were administered in spring of Grade 2. Together, MAP and STEP scores explained 65 to 72% of the variance in Cohort 1 and 51 to 65% of the variance in Cohort 2, when predicting scores on the Georgia Milestones.

Classification accuracy of STEP and MAP

In both cohorts, a large proportion of students did not meet proficiency on the state assessment (Cohort 1 = 66% [*n* = 135]; Cohort 2 = 57% [*n* = 66]). As shown in Table 3.3, classification accuracy results fell in the good or acceptable range when using statistically-optimized cut-scores for MAP, across cohort and season. In Cohort 1, this cut-score was 199 for spring of second grade, 196 for fall of third grade, and 206 for spring of third grade – all higher than the national norms. Likewise, with Cohort 2, the optimized cut-score was 199 in spring of second grade, 197 in fall of third grade, and 204 in spring of third grade; also higher than the national norms. Cohort 1 spring second- and third-grade MAP cut-scores resulted in the highest AUC values of .931 and .930 respectively, compared to MAP cut-scores in the fall of third grade at .899. AUC values were similar in Cohort 2, albeit slightly lower. However, all AUC confidence intervals overlapped. When examining sensitivity, specificity, PPV and NPV between optimized publisher-recommended cut scores, large differences were observed. In general, sensitivity and NPV were lower and specificity and PPV were higher with the publisher-provided benchmarks. Regarding STEP, NPV for the STEP assessment approached 80% in Cohort 1 with NPV in spring second-grade administration equaling 79%. In Cohort 2, spring second-grade STEP scores resulted in the only NPV above 80%.

Table 3.1

Descriptive Statistics and Intercorrelations Among Study Variables

			Correlations						
Variable	*n*	*M*	1	2	3	4	5	6	7
Cohort 1									
Spring 2									
1. MAP	197	191.32	–						
2. STEP	196	8.18	.76*	–					
Fall 3									
3. MAP	99	188.11	.75*	.73*	–				
4. STEP	208	8.65	.76*	.92*	.71*	–			
Spring 3									
5. MAP	206	199.23	.83*	.73*	.78*	.72*	–		
6. STEP	207	9.82	.74*	.79*	.75*	.85*	.75*	–	
7. GM	209	504.05	.82*	.78*	.75*	.77*	.82*	.78*	–
Cohort 2									
Spring 2									
1. MAP	108	195.62	–						
2. STEP	108	9.04	.61*	–					
Fall 3									
3. MAP	114	193.28	.76*	.64*	–				
4. STEP	115	9.32	.68*	.87*	.67*	–			
Spring 3									
5. MAP	115	200.88	.71*	.58*	.83*	.66*	–		
6. STEP	115	10.22	.68*	.82*	.68*	.89*	.65*	–	
7. GM	115	522.13	.67*	.62*	.77*	.68*	.77*	.71*	–

Note. MAP = Measures of Academic Progress; STEP = Strategic Teaching and Evaluation of Progress; GM = Georgia Milestones.

* $p < 0.01$.

Table 3.2

Summary of Hierarchical Regression Analyses Using MAP and STEP to Predict Georgia Milestones Performance

Predictor	R^2	ΔR^2
Cohort 1		
Spring 2 MAP	.65*	
Spring 2 MAP, Spring 2 STEP	.72*	.07*
Fall 3 MAP	.52*	
Fall 3 MAP, Fall 3 STEP	.65*	.13*
Spring 3 MAP	.65*	
Spring 3 MAP, Spring 3 STEP	.72*	.07*
Cohort 2		
Spring 2 MAP	.43*	
Spring 2 MAP, Spring 2 STEP	.51*	.08*
Fall 3 MAP	.56*	
Fall 3 MAP, Fall 3 STEP	.61*	.04*
Spring 3 MAP	.59*	
Spring 3 MAP, Spring 3 STEP	.65*	.06*

Note. MAP = Measures of Academic Progress; STEP = Strategic Teaching and Evaluation of Progress.

*$p < .001$.

Table 3.3

Classification Accuracy of Screening Measures to Predict Failure on the Georgia Milestones

Screening Measure	Cut-score	AUC	*SE*	95% CI	Sens (%)	Spec (%)	PPV (%)	NPV (%)
Cohort 1								
MAP								
Spring 2	199	.931	.017	[.897, .965]	89	79	89	78
	188.7[a]	–	–	–	60	97	98	56
Fall 3	196	.899	.034	[.821, .966]	89	77	88	79
	188.3[a]	–	–	–	67	91	93	60
Spring 3	206	.930	.017	[896, .964]	92	78	89	83
	198.6[a]	–	–	–	69	94	96	61
STEP								
Spring 2	9	–	–	–	89	71	84	79
Spring 3	12	–	–	–	87	81	91	73
Cohort 2								
MAP								
Spring 2	199	.894	.029	[.837, .951]	85	79	84	80
	188.7[a]	–	–	–	49	100	100	60
Fall 3	197	.903	.029	[.846, .959]	85	82	86	80
	188.3[a]	–	–	–	53	96	94	62
Spring 3	204	.912	.027	[.860, .964]	86	82	86	82
	198.6[a]	–	–	–	65	96	96	67
STEP								
Spring 2	9	–	–	–	86	63	62	87
Spring 3	12	–	–	–	81	83	89	71

Note. MAP = Measures of Academic Progress; STEP = Strategic Teaching and Evaluation of Progress; AUC = area under the curve; PPV = positive predictive value; NPV = negative predictive value. [a] = benchmark provided by MAP

CHAPTER 4

DISCUSSION

Throughout the nation, educators use screening tools like MAP and STEP to predict students' reading proficiency, as measured by end-of-year standardized state assessments. Early remediation is critical for students who may be at risk in reading, as mandated tests are increasingly used to help determine grade promotion in many states starting in Grade 3 (Weyer, 2018). Schools use MAP and STEP simultaneously as screening measures; however, limited empirical evidence exists to support the administration of both assessments. Research supports curriculum-based measures of reading for screening purposes (Reschly et al., 2009) yet, little to no published research exists to support the technical adequacy and classification accuracy of MAP and STEP. To address these gaps in the literature, both the criterion validity and classification accuracy of MAP and STEP were evaluated. The present study investigated the following: (1) the concurrent and predictive validity of MAP and STEP with the high-stakes test in third grade, (2) the incremental validity of administering STEP in addition to MAP, (3) and the classification accuracy of statistically-optimized cut-scores for MAP compared to publisher recommended cut-scores for MAP and publisher suggested benchmarks for STEP. The conditional probability of student outcomes was used to report sensitivity, specificity, NPV, and PPV.

With regard to criterion validity, results from the correlational analysis indicated that MAP and STEP were statistically significant predictors of the third-grade state test. Overall, MAP correlations were higher when compared to STEP at all but one time point (i.e., fall third grade in Cohort 1). Additionally, correlation coefficients in Cohort 2 were generally lower than Cohort 1 for both screening measures, which may be due to statistically significant cohort differences assessed prior to the correlational analysis. Scores in the first cohort indicated stronger relations for both screeners in the spring of each grade, suggesting better prediction of student performance at the end of each school year. Furthermore, in Cohort 2, there were stronger associations for both screeners in third grade across the fall and spring, indicating better prediction of student performance within the same year of state assessment administration. These findings offer initial evidence to support STEP as an adequate representation of reading proficiency, and stronger evidence for MAP as an even better indicator of reading skills. MAP may be a better tool for screening; however, more consistent evidence is needed to support reliability given the emerging literature.

Strong associations between MAP and the state assessment corroborate previous research of strong associations between MAP and other standardized assessments of reading achievement (Ball & O'Connor, 2016; January & Ardoin, 2015; Klingbeil et al., 2015, 2017). However, the concurrent relation between MAP and the standardized assessment was slightly lower than findings from the one previous study that evaluated concurrent validity with a norm-referenced assessment (January & Ardoin, 2015). This slight difference in findings may be due to a number of reasons including sample characteristics, instructional practices, or characteristics of the outcome measure.

Generally, the findings from the current study were similar to or higher than the range of correlations reported in previous studies for predictive validity using either MAP or state assessments as the criterion measure (Ball & O'Connor, 2016; Klingbeil et al., 2015; 2017). Additionally, the recent study conducted by MAP test publishers corresponded precisely with the concurrent validity findings from the first cohort in the present study (NWEA, 2016). Regarding STEP, no known studies exist that investigate its technical adequacy aside from the unpublished technical report authored by the test developers. The correlations reported in Cohort 1 exceeded associations reported in the technical report and coefficients in Cohort 2 were similar to those found in the technical report (Kerbow & Bryk, 2005).

To investigate the predictive value of both assessments, MAP and STEP scores for each season were entered to measure the incremental validity of a combined screening system. MAP was entered into each model first because there is more research to support its validity as compared to STEP. For each timepoint, statistical models with both MAP and STEP demonstrated statistically larger associations than MAP alone. STEP added statistically significant predictive value when evaluated in combination with MAP to predict performance on the state assessment. This finding may suggest that it is beneficial for educators to administer STEP in addition to MAP, for the purpose of increasing predictive utility of student performance. This could be helpful during seasons or grades when MAP alone demonstrates lower correlations with the GA Milestones compared to other periods of administration. However, educators should also consider the practical significance of administering STEP in addition to MAP given the negligible added predictive value and only slight increase in explained variance for each time point.

Furthermore, educators must consider the time and resources needed to administer STEP to students. It may be that educators find other uses for STEP data, such as informing instruction. Nonetheless, it is important to consider the purposes of assessment when selecting and administering measures.

In the current study, statistically-optimized MAP cut-scores were higher than publisher-recommended benchmarks. This finding may support the investigation of cut-scores that are tailored to the specific sample of interest, for increased classification accuracy. Additionally, despite significant differences in achievement across cohorts, statistically-optimized MAP cut-scores exhibited a better balance between sensitivity and specificity relative to STEP and publisher-recommended benchmarks. Sensitivity and specificity for both MAP and STEP varied across grade and season, yet they approached acceptable levels of 90% sensitivity and 80% specificity. Additionally, MAP had good overall classification accuracy, with AUC values either just below or above .90, which was generally higher than the report from the technical manual (NWEA, 2011). Both MAP and the state test are group-administered computer assessments which may explain better predictive accuracy in comparison to STEP, which is an individual student-teacher conference measure of reading. STEP did, however, demonstrate moderate to strong predictive accuracy across cohorts.

Overall, test developer STEP benchmarks yielded accurate predictions of students who failed the end-of-year assessment in third grade. Accurate prediction of students who were not at risk requires further investigation, given the relatively lower specificity in Cohort 2 for spring of third grade. Additionally, reports for NPV across grade and season were inconsistent with the 80% to 86% that test developers reported (Kerbow & Bryk, 2005). With one exception (spring Grade 2, Cohort 2), findings either approached but did not meet the NPV reported for percentage of students expected to perform proficiently on the end-of-year state assessment.

4.1 LIMITATIONS

Despite the moderate to strong evidence in support of the technical adequacy of MAP and STEP, there are potential limitations of the current study. First, given the homogenous sample of students with respect to race, ethnicity, and socioeconomic status across cohorts, the generalizability of findings is restricted. However, this also represents a strength of this study, as limited published research in this area exists that includes racial/ethnic minority students. A second limitation was that there was a high percentage of students in each cohort who did not pass the state assessment was high across the schools where data was collected. The issue of high base rates might require a more tailored approach of statistical analysis to account for inflated results. Third, we were unable to directly assess fidelity of administration of the STEP tool, given the use of archival data. However, this also reflects data that were collected under typical conditions in the participating schools. Further, teachers were trained extensively on administration of STEP. Fourth, the percentage of missing MAP data were higher than desired in the current study. By using a pairwise present approach instead of listwise deletion, we preserved as much information as possible and reduced bias. Moreover, controlling for confounding variables (e.g., different school districts or classrooms) might alter findings.

Although STEP demonstrated adequate criterion validity and partially sufficient classification accuracy, optimal cut-scores were not created for this measure due to the categorical nature of the STEP levels. It may have been unreasonable to compare test developer STEP benchmarks to locally-derived MAP cut-scores that were optimal for the sample of students in this study. Moreover, fall benchmarks were not provided in the STEP technical report to accurately investigate validity during the beginning of school year for third grade. As a final limitation, the testing windows for the STEP assessment were too large in the winter to be compared to the winter administration for MAP, thus, only fall and spring administration scores were analyzed.

4.2 FUTURE DIRECTIONS

There are also several potential future research directions. For example, future research is needed to replicate and extend findings from the current study using different populations and criterion assessments. Further research may also continue to investigate the classification accuracy of a combined system of screening since STEP scores added some incremental validity to MAP scores. The continued evaluation of validity for MAP and STEP will add to a growing literature base while also helping to inform the continued concurrent use of both measures with the ultimate goal of optimizing student success.

4.3 IMPLICATIONS FOR PRACTICE

Findings from the current study indicate that MAP and STEP are adequate predictive measures of reading proficiency. Both screeners can be used to adequately identify whether students are at risk for poor performance on the state assessment in the spring of third grade. Educators are encouraged, nonetheless, to consider the contextual utility of dual administration since both measures are costly and time intensive to use. MAP has sufficient technical adequacy to be used in isolation, which may be informative for schools wanting to conserve time and resources. Alternatively, for the purpose of maximizing prediction of student performance and potentially informing instruction, findings from the current study support only a slight increase in predictive value of administering STEP in addition to MAP. Future research, however, will determine whether classification accuracy of a multivariate system of screening using both STEP and MAP exceeds single measure utility. Ultimately educators should elect to adopt a screening system that promotes the provision of timely and appropriate intervention to as many students at risk for reading failure. Convenience of administration compared to other quick and already supported measures, acceptability by those administering the screeners, and the context in which screening data is used should all help guide the utility of STEP and MAP in schools. Additionally, educators should consider using locally derived cut-scores as opposed to national normed benchmarks, if that is more appropriate for their school. Regardless of the methods that educators use to identify students at risk, schools will benefit from a valid system of screening that helps inform appropriate and effective delivery of tailored intervention for students struggling to meet reading proficiency standards.

REFERENCES

Aiken, L. R. (1985). *Psychological testing and assessment.* Needham Heights, MA: Allyn & Bacon.

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

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

Bennett, I. M., Frasso, R., Bellamy, S. L., Wortham, S., & Gross, K. S. (2013). Pre-teen literacy and subsequent teenage childbearing in a US population. *Contraception*, *87*(4), 459–464. doi.org/10.1016/j.contraception.2012.08.020

Burns, M. K., Haegele, K., Petersen-Brown, S. (2014). Screening for early reading skills: using data to guide resources and instruction. 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. 171–197). Washington, DC: American Psychological Association.

Catts, H. W., Nielsen, D., Bridges, M., Liu, S., & Bontempo, D. E. (2015). Early identification of reading disabilities within a RTI framework. *Journal of Learning Disabilities*, *48*(3), 281–297. doi.org/10.1177/0022219413498115

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). Washington, DC: American Psychological Association.

Darney, D., Reinke, W. M., Herman, K. C., Stormont, M., & Ialongo, N. S. (2013). Children with co-occurring academic and behavior problems in first grade: Distal outcomes in twelfth grade. *Journal of School Psychology, 51*, 117-128. doi.org/10.1016/j.jsp.2012.09.005

Daniel, S. S., Walsh, A. K., Goldston, D. B., Arnold, E. M., Reboussin, B. A., & Wood, F. B. (2006). Suicidality, school dropout, and reading problems among adolescents. *Journal of Learning Disabilities*, *39*(6), 507–514. doi.org/10.1177/00222194060390060301

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*(3), 219–232. doi.org/10.1177/001440298505200303

Every Student Succeeds Act of 2015, Pub. L. No. 114-95 (2015).

Georgia Department of Education (2015).

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

Individuals with Disabilities Education Improvement Act of 2004, U.S.C. H.R 1350 (2004).

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. doi.org/10.1177/1534508415579095

January, S.-A. A., Ardoin, S. P., Christ, T. J., Eckert, T. L., & White, M. J. (2016). Evaluating the interpretations and use of curriculum-based measurement in reading and word lists for universal screening in first and second grade. *School Psychology Review*, *45*(3), 310–326. doi.org/10.17105/SPR45-3.310-326

January, S.-A. A., Mason, W. A., Savolainen, J., Solomon, S., Chmelka, M. B., Miettunen, J., . . . Järvelin, M.-R. (2017). Longitudinal pathways from cumulative contextual risk at birth to school functioning in adolescence: Analysis of mediation effects and gender moderation. *Journal of Youth and Adolescence, 46*, 180-196. doi.org/10.1007/s10964-016-0560-9

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*, 582–600.

Jenkins, J. R. & Johnson, E. S. (2009) Universal screening for reading problems: Why and how should we do this? Retrieved from http://www.rtinetwork.org/Essential/Assessment/Universal/ar/ReadingProblems

Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (Eds.). (2016). *Handbook of Response to Intervention: The Science and Practice of Multi-Tiered Systems of Support* (2nd ed.). New York, NY: Springer Science+Business Media. doi.org/10.1007/978-1-4899-7568-3

Kerbow, D. & Bryk, A. (2005). *STEP literacy assessment: Technical report of validity and reliability.* Retrieved from www.ibrarian.net/.../paper/STEP_Literacy_Assessment_Technical_Report_of_Vali.pdf

Kettler, R. J., Glover, T. A., Albers, C. A., & Feeney-Kettler, K. A. (2014). *Universal screening in educational settings: Evidence-based decision making for schools.* Washington, DC: American Psychological Association.

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. doi.org/10.1016/j.jsp.2014.06.002

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

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. doi.org/10.1177/0741932517697446.21839

McGill-Franzen, A. (1987). Failure to learn to read: Formulating a policy problem. *Reading Research Quarterly, 22*(4), 475-490. doi:10.2307/747703

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. doi.org/10.3102/0013189X018002005

Nelson, J. M. (2009). Psychometric properties of the Texas primary reading inventory for early reading screening in kindergarten. *Assessment for Effective Intervention, 35*(1), 45–53. doi.org/10.1177/1534508408326205

Nelson, P. M., Van Norman, E. R., & Lackner S. K. (2017). A comparison of methods to screen middle school students for reading and math difficulties students for reading and math difficulties, (March).

Northwest Evaluation Association. (2011). *Technical manual for measures of academic progress and measures of academic progress for primary grades*. Retrieved from: https://www.richland2.org/.../5-1-NWEA-Technical-Manual-for-MAP-and-MPG.pdf

Northwest Evaluation Association. (2015). *NWEA measures of academic progress interim assessments for grades k – 12.* Retrieved from https://in.lcms.org/wp-content/uploads/2016/09/NWEA-Overview-of-Services.pdf

Northwest Evaluation Association. (2016).

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-Based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology*, *47*(6), 427–469. doi.org/10.1016/J.JSP.2009.07.001

Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly*, *42*(4), 546–567. doi.org/10.1598/RRQ.42.4.5

Urban Education Institute, University of Chicago (2011). *STEP*. Retrieved from https://uchicagoimpact.org/tools-training/step/

U.S. Department of Health and Human Services, National Institute of Child Health and Human Development. (2000). Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups (NIH Publication No. 00–4754). Retrieved from https://www.nichd.nih.gov/publications/pubs/nrp/Documents/report.pdf

Vitaro, F., Brendgen, M., Larose, S., & Trembaly, R. E. (2005). Kindergarten disruptive behaviors, protective factors, and educational achievement by early adulthood. *Journal of Educational Psychology, 97*(4), 617-629. doi.org/10.1037/0022-0663.97.4.617

Vaughn, S., Fletcher, J. M., Francis, D. J., Denton, C. A., Wanzek, J., Wexler, J., … Romain, M. A. (2008). Response to intervention with older students with reading difficulties. *Learning and Individual Differences*, *18*(3), 338–345. doi.org/10.1016/J.LINDIF.2008.05.001

Walker, H. M., Small, J. W., Severson, H. H. Seeley, & Feil, E. G. (2014). Multiple-gating approaches in universal screening within school and community settings. 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. 47–75). Washington, DC: American Psychological Association.

Wayman, M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41*(2), 85–120. doi.org/10.1177/00224669070410020401

Weyer, M. (2018). *A look at third-grade reading retention policies.* Retrieved from http://www.ncsl.org/documents/legisbriefs/2018/june/LBJune2018_A_Look_at_Third_Grade_Reading_Retention_Policies_goID32459.pdf

Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education*, *31*(6), 412–422. doi.org/10.1177/07419325083274