Examining the Relationship Between a Universal Screening Measure and a State Reading Assessment

Karen G. Smith
University of Maine, karen.g.smith@maine.edu

Follow this and additional works at: https://digitalcommons.library.umaine.edu/etd

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation
Smith, Karen G., 'Examining the Relationship Between a Universal Screening Measure and a State Reading Assessment' (2019). Electronic Theses and Dissertations. 3060.
https://digitalcommons.library.umaine.edu/etd/3060

This Open-Access Thesis is brought to you for free and open access by DigitalCommons@UMaine. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of DigitalCommons@UMaine. For more information, please contact um.library.technical.services@maine.edu.
EXAMINING THE RELATIONSHIP BETWEEN A UNIVERSAL SCREENING MEASURE AND A STATE READING ASSESSMENT FOR MIDDLE-LEVEL STUDENTS

By

Karen G. Smith

B.S.N. University of Southern Maine, 1994
M.Ed. University of Maine, 2007

A Dissertation
Submitted in Partial Fulfillment of the Requirements for the Degree of Doctorate in Education

The Graduate School
The University of Maine
August 2019

Advisory Committee:

James Artesani, Associate Dean of Accreditation and Graduate Affairs; Associate Professor of Special Education, Co-Advisor
Sara Flanagan, Assistant Professor of Special Education, Co Advisor
Brian Cavanaugh, Assistant Professor of Special Education
Craig Mason, Professor of Education & Applied Quantitative Methods
© 2019 Karen G. Smith

All Rights Reserved
EXAMINING THE RELATIONSHIP BETWEEN A UNIVERSAL SCREENING MEASURE AND A STATE READING ASSESSMENT FOR MIDDLE-LEVEL STUDENTS

By

Karen G. Smith

Dissertation Co-Advisors: Dr. James Artesani and Dr. Sara Flanagan

An Abstract of the Dissertation Presented in Partial Fulfillment of the Requirements for the Degree of Doctor in Education August 2019

Many adolescents struggle with reading comprehension, despite the emphasis on reading instruction in recent decades. Evidence suggests that informational text is particularly challenging for students. Universal screening data can be useful when identifying students who are at risk of not meeting proficiency standards on high-stakes assessments and in need of reading intervention. To implement assessments within a multi-tiered framework, schools must have psychometrically adequate tools.

Using data of 473 students in Grades 6 through 8 from two Maine middle schools, this study examined the relationship between the NWEA’s Measures of Academic Progress in Reading (MAP-R) and Maine’s eMPowerME English Language Arts/Literacy (ELA/L) test by Measured Progress. Logistic regression models were statistically significant, with MAP scores explaining 54.6% and 58.2% of the variance in proficiency on the eMPowerME and correctly classifying 83% and 80% of cases for Grades 6-7 and 7-8. Gender, SES, and disability status were used to determine if value was added by combining MAP scores with student
characteristics. The added demographic variables were less robust predictors of reading achievement when combined with MAP scores compared to MAP scores alone.

This investigation also determined the minimum MAP scores needed in spring of Grade 6 and 7 demonstrating a student was on track to meet proficiency standards on eMPowerME one year later. MAP scores resulted in high AUC values; however, diagnostic accuracy was below the acceptable level recommended for a screener when using NWEA-provided cut scores. By using locally derived cut scores, the diagnostic accuracy was improved by maximizing sensitivity and specificity to an acceptable level.

A secondary purpose of this study was to determine whether differences existed between MAP subtests scores and to examine the unique contribution of MAP subtests to the eMPowerME. Students performed significantly higher on the vocabulary acquisition and use (VAU) subtest compared to either the literary or informational text subtests. There was not a statistically significant difference between literary and informational text scores. The VAU subtest had the weakest correlations with eMPowerME. MAP literary text for Grade 6-7 and informational text for Grade 7-8 accounted for the highest degree of variance in eMPowerME scores.
DEDICATION

It is my genuine gratefulness and warmest regard that I dedicate this dissertation work to Dr. Raymond and (the late) Matiana Glass. Both worked tirelessly over their careers to advocate for children and youth with academic and social/emotional disabilities and their families. Ray and Matiana instilled in me the value of evidence-based practices and using data to inform decision-making. Through their example, they taught me the core principles and ethical standards about what it means to be a special education professional that has guided me in preparing UMF pre-service teachers, providing support to in-service professionals, and engaging with scholarly activities. It is without a doubt that Ray and Matiana were a driving force that put me on the path of where I am today.
ACKNOWLEDGEMENT

I want to give a sincere thank you to all my past and current committee members (Dr. James Artesani—Co-Chair, Dr. Sara Flanagan—Co-Chair, Dr. Brian Cavanaugh, Dr. Craig Mason, Dr. Eric Pandiscio, Dr. Janet Spector, Dr. Theodore Coladarci, and Dr. Brian Doore). The support and guidance given to me by these individuals throughout this process are immeasurable. Also, I could not have completed this work without the loving support from family. I truly am indebted to my husband, Joel, and two sons, Samuel and Joshua, for their tremendous patience throughout this entire journey.
# TABLE OF CONTENTS

DEDICATION ......................................................................................................................... iii

ACKNOWLEDGMENT ............................................................................................................. iv

LIST OF TABLES .................................................................................................................... ix

LIST OF FIGURES ................................................................................................................ x

Chapter

1. INTRODUCTION ............................................................................................................... 1

   Background and Statement of the Problem ................................................................. 2

   Purpose of the Study and Research Questions ......................................................... 12

   Significance of the Study ............................................................................................. 14

   Organization of the Study ............................................................................................ 15

2. REVIEW OF THE LITERATURE ....................................................................................... 16

   Historical and Current Focus on Informational Text ................................................ 18

      Informational and Literary Text Characteristics ................................................... 19

      Comparing the Comprehension of Informational Versus Literary Text .............. 23

      Summary of Informational Text .............................................................................. 30

   Technical Adequacy of Screening Assessments, Procedures, and
   Decision Making .............................................................................................................. 31

      Number of Screeners and Routes to Intervention ................................................. 36

      General Outcome Measures (GOMs) .................................................................... 39

      Computer Adaptive Tests (CATs) .......................................................................... 43

      Summary of Universal Screening .......................................................................... 44
3. METHOD ......................................................................................................................... 46
   Procedure for Data Collection ......................................................................................... 47
   Sample Description ........................................................................................................ 49
   Measures .......................................................................................................................... 51
      Measures of Academic Progress-Reading (MAP-R) ..................................................... 51
      eMPowerME English Language Arts/Literacy (ELA/L) ............................................... 53
   Procedure for Data Analysis .......................................................................................... 56
   Summary .......................................................................................................................... 58

4. RESULTS ....................................................................................................................... 59
   Descriptive Statistics ...................................................................................................... 61
      Description of the Sample ............................................................................................. 61
         Participation in MAP-R by Grade Cohort ................................................................. 62
         Participation in MAP-R by Student Demographics ................................................. 62
         Participation in MAP-R by eMPowerME ELA/L
            Achievement Levels ............................................................................................... 63
            Performance on MAP-R and eMPowerME ELA/L ............................................. 65
   Correlation Analysis ...................................................................................................... 69
   Logistic Regression Analysis ......................................................................................... 72
   ROC Curve Analysis ...................................................................................................... 77
   ANOVA and t-Test Analysis ........................................................................................... 81
      Differences Between MAP-R Subtests ..................................................................... 82
         MAP-R Spring 2016 ................................................................................................. 82
         MAP-R Spring 2017 ............................................................................................... 83
Differences within MAP-R Subtests by Student Demographics ....................... 84

Literary Text Comprehension – Gender ......................................................... 85
Literary Text Comprehension – Socioeconomic Status (SES) ......................... 85
Literary Text Comprehension – Disability Status ........................................... 86
Informational Text Comprehension – Gender ................................................ 86
Informational Text Comprehension – Socioeconomic Status (SES) ............... 87
Informational Text Comprehension – Disability Status .................................. 87
Vocabulary Acquisition and Use – Gender ...................................................... 88
Vocabulary Acquisition and Use – Socioeconomic Status (SES) ................. 89
Vocabulary Acquisition and Use – Disability Status ..................................... 89
Hierarchical Regression Analysis .................................................................. 90
Summary .......................................................................................................... 92

5. DISCUSSION ................................................................................................. 98

A Review of Data Analyses and Results ......................................................... 98

Concurrent and Predictive Validity of MAP-R with eMPowerME ELA/L ................................................................. 99

MAP-R with Demographic Variables to Predict Performance on eMPowerME ELA/L ................................................................. 102

Optimizing MAP-R Cut Scores in Predicting Performance on eMPowerME ELA/L ................................................................. 103

MAP-R Performance Between and Within Subtests .................................... 105

MAP-R Subtests as Components of eMPowerME ELA/L ............................. 107

Implications for Practitioners and Policy Makers ........................................ 107
Assumptions, Limitations, and Future Research ........................................ 110
Summary ........................................................................................................ 112
REFERENCES ................................................................................................. 114
BIOGRAPHY OF THE AUTHOR ................................................................. 142
LIST OF TABLES

Table 1. School Demographic Data (Grades 6, 7, 8) ........................................51
Table 2. eMPowerME English Language Arts/Literacy Achievement Level
  Descriptors and Score Ranges .................................................................55
Table 3. MAP-R Participation Rates by Grade Cohort ....................................64
Table 4. MAP-R Participation Rates by Student Demographics .....................64
Table 5. MAP-R Participation Rates by eMPowerME ELA/L
  Achievement Levels ............................................................................64
Table 6. MAP-R Participation by Grade Cohort According to
  eMPowerME ELA/L Achievement Levels ..............................................66
Table 7. Descriptive Statistics for MAP-R Total and Subtests
  and eMPowerME-ELA/L ......................................................................68
Table 8. Descriptive Statistics eMPowerME Spring 2017
  by Student Demographics ..................................................................69
Table 9. Correlations Matrix of MAP–R and eMPowerME ELA/L ..................72
Table 10. Logistic Regression of Spring 2016 MAP-R and Spring 2017
  eMPowerME ELA/L ............................................................................77
Table 11. MAP-R Cut Scores, Sensitivity, and Specificity Associated
  with Percentiles ....................................................................................80
Table 12. Summary of Hierarchical Regression Analyses Using Spring 2016
  MAP-R Goal Areas to Predict Performance on Spring 2017
  eMPowerME ELA/L ............................................................................91
Table 13. Summary of Research Questions, Data Analyses, and Results ...........................96
LIST OF FIGURES

Figure 1. Roc Curves Spring 2016 MAP-R and Spring 2017
eMpowerME ELA/L .................................................................79
Chapter 1

INTRODUCTION

Recent data from the National Assessment of Educational Progress (NAEP) indicate that many adolescents in the United States struggle with reading comprehension (Edmonds et al., 2009; Flagella-Luby, Graner, Deshler, & Drew, 2012; Morsy, Kieffer, & Snow, 2010; National Center for Education Statistics, 2013, 2015, 2019; Scammacca et al., 2007). In response to the growing evidence that suggests understanding informational text, in particular, is challenging for many students (Best, Floyd, & McNamara, 2008; Cummins, 2013; Dennis, 2013; Eason, Goldberg, Young, Geist, & Cutting, 2012; McNamara, Ozuru, & Floyd, 2011; Sanacore & Palumbo, 2009; Thompson et al., 2012; Zabrucky & Ratner, 1992), national organizations, content experts, leading researchers, and assessment developers now place an increased emphasis on this text type (Duke, 2000; Jeong, Gaffney, & Choi, 2010; Kamil & Lane, 1997; National Assessment Governing Board [NAGB], 2015; National Governors Association [NGA] Center for Best Practices & Council of Chief State School Officers [CCSSO], 2010a). The ability to provide effective reading intervention, however, relies on the use of efficient and accurate universal screening measures. Purposes of universal screening assessments include identifying those students who are at risk of reading failure (Mellard, McKnight, & Woods, 2009) and predicting student performance on high-stakes reading achievement tests (Hintze & Silberglitt, 2005). It is critical, therefore, that education practitioners make use of reading assessment data to identify students who have difficulty with text comprehension and apply the information available to inform practice.

This study investigated the relationship between a universal screening measure, the Northwest Evaluation Association’s Measures of Academic Progress in Reading (NWEA MAP-
R), and Maine’s state summative assessment, the eMPowerME English Language Arts/Literacy (ELA/L) test by Measured Progress. The primary purpose of the study was to examine the concurrent and predictive validity between these two measures (MAP-R and eMPowerME ELA/L) for students in Grades 6-7 and 7-8. Gender, socioeconomic status, and disability status were used to ascertain if any value was added by using the MAP-R scores along with certain student characteristics in predicting performance on the eMPowerME ELA/L. This investigation also determined the minimum MAP-R scores needed in the spring of Grade 6 and 7 that demonstrated a student was on track to meet proficiency standards on the eMPowerME one year later in Grade 7 and 8. The results are intended to provide critical information for education practitioners when using the MAP-R as a universal screening measure when identifying students who are at risk of reading difficulties. The secondary purpose of the study was to examine whether there were significant differences between middle-level students' performance on the MAP-R subtests (i.e., literary text, informational text, vocabulary acquisition and usage). In doing so, the study was designed to contribute to the literature on the efficacy of using reading subscores to identify the needs of adolescent readers differences when comprehending literary and informational passages.

**Background and Statement of the Problem**

Reading comprehension is a critical foundation for learning in school (Chall, 1996; Snow, 2002) and it is an essential component of college and workforce readiness (Achieve, 2004; ACT, 2004, 2006a, 2006b, 2008; Wyatt, Wiley, Camara, & Proestler, 2011). For decades, literacy experts have articulated concerns about the reading performance of students in fourth through twelfth grades and in particular, the gap in achievement among certain groups of students. For example, White students have consistently outperformed their black and Hispanic
peers on the NAEP reading test (Bohrnstedt, Kitmitto, Ogut, Sherman, & Chan, 2015; Hemphill & Vanneman, 2011). In a recent meta-analysis, Gilmour, Fuchs, and Wehby (2019) determined that students with disabilities performed 1.17 standard deviations or 3.3 years below their typically developing peers in reading. The gap in literacy achievement between students from economically advantaged and disadvantaged backgrounds also is well documented in the literature. In a recent analysis of long-term trends in national and international test data, Hanushek, Peterson, Talpey, and Woessmann (2019) learned that the disparities are large and unwavering over the past half-century. That is, by Grade 8, students in the 90th percentile of socioeconomic status (SES) distribution are consistently three to four years ahead than those at the 10th percentile.

In addition to gaps in race/ethnicity, disability status, and SES, numerous national studies found that, on average, females outperform males on English Language Arts (ELA) tests in the U.S. (e.g., Chatterji, 2006; Robinson & Lubienski, 2011) and these literacy achievement gaps vary from state to state (Pope & Sydnor, 2010). In particular, recent investigations indicated that community and family socioeconomic contexts differentially affect male and female academic outcomes and educational attainment (Autor, Figlio, Karbownik, Roth, & Wasserman, 2016; Chetty, Hendren, Lin, Majerovitz, & Scuderi, 2016). The reported literature showed that gaps in ELA scores favored kindergarten female performance over male performance by 0.15 and 0.20 standard deviations (Chatterji, 2006; Husain & Millimet, 2009; Robinson & Lubienski, 2011). According to Robinson and Lubienski (2011), the gender gap narrows moderately through fifth grade but widens again by the eighth grade. Moreover, the literacy gap between females and males has remained somewhat constant over the last several decades (Fahle & Reardon, 2018).
The concerns about overall literacy attainment and the achievement gaps among certain groups of students have prompted the development of numerous task force reports and policy documents (e.g., ACT, 2008; Biancarosa & Snow, 2006; Boardman et al., 2008; Carnegie Council on Advancing Adolescent Literacy, 2010; Center on Education Policy, 2007; Fisher & Ivey, 2006; Heller & Greenleaf, 2007; International Reading Association, 2002), and they have contributed to legislative mandates aimed at increasing student reading achievement (e.g., Every Student Succeeds Act [ESSA] of 2015; Individuals with Disabilities Education Improvement Act [IDEIA] of 2004; No Child Left Behind [NCLB] Act of 2002). In concert with standards-driven reading instruction, current educational models emphasize the importance of using assessment data for school accountability purposes (Rupp & Lesaux, 2006) and to inform instructional decision-making (Mandinach, 2012). Indeed, the latest reauthorization of the Elementary and Secondary Education Act (ESEA) of 1965, ESSA of 2015 continues previous accountability provisions under NCLB of 2002. NCLB mandated that states test students in reading and math proficiency in Grades 3 through 8 and once in high school. Informal reading assessments have always been a part of classroom instruction, but current practices reflect an increase in the number of standardized tests given. In turn, the tests are tied to improving student reading performance on high-stakes state assessments. An issue educators must grapple with is finding the balance between the allocated time spent on assessments with the necessary time needed to provide high-quality instruction and intervention (VanDerHeyden & Harvey, 2013). In addition to navigating when and how to assess, many practitioners have an inadequate understanding of how to sufficiently use assessment information to guide changes in instruction or programming (Ball & Christ, 2012).
Despite efforts made toward addressing the adolescent literacy problem, little is known about the characteristics of struggling adolescent readers beyond demographic differences (Speece et al., 2010). As has been the case with younger students, profile research has shown that adolescents who struggle with reading may have difficulties with decoding skills, vocabulary knowledge, fluency, or reading comprehension or a combination of factors (Brasseur-Hock, Hock, Kieffer, Biancarosa, & Deshler, 2011; Floyd, Bergeron, & Alfonso, 2006; Snow & Biancarosa, 2003). In a 2000 report by the National Reading Panel, experts identified the five pillars of reading instruction including phonemic awareness, phonics, oral reading fluency, vocabulary, and comprehension. Each of these areas has been linked as a predictor of reading achievement and are viewed as a hierarchy of reading instruction (National Institute of Child Health and Human Development [NICHD], 2000).

Some researchers believe that the emphasis on high-stakes assessments has led to the narrowing of the curriculum to focus on constrained skills (Linn, 2000; Rupp & Lesaux, 2006). According to Paris (2005), constrained skills have a ceiling for mastery, such as phonemic awareness and letter knowledge, while unconstrained skills are those without a ceiling for mastery like vocabulary and comprehension. Paris argues that constrained skills are more straightforward to assess than unconstrained skills. Hence, much research has focused on the instruction and assessment of constrained skills and less research on vocabulary and comprehension, signifying a higher level of importance on constrained skills. Subsequently, less instructional time is devoted to the multifaceted process of vocabulary and comprehension (Lesaux & Kieffer, 2010).

Reading comprehension is a complex, multidimensional process that requires coordination and integration of multiple underlying processes (Kintsch, 1988, 2013; Perfetti,
Effective reading comprehension, therefore, relies on a multitude of factors that relate to the text, the reader, and the act of comprehension itself (Dennis, 2013; Rupp & Lesaux, 2006). As such, researchers have found that the influence of component skills for successful reading comprehension varies across readers and changes as texts become more complex indicating that readers are heterogeneous with diverse instructional needs (Carlisle, 2000; Dennis, 2013; Buly, & Valencia, 2002). Reading experts have reported that adolescents that struggle with reading are mostly capable of the multifaceted process of reading but require intensive instruction using texts that support their development of sophisticated vocabulary and comprehension skills (Dennis, 2013; Lesaux & Kieffer, 2010; Rupp & Lesaux, 2006).

Several studies have suggested that students’ reading performance differs according to text type (e.g., Eason et al., 2012; Stamboltzis & Pumfrey, 2000). In particular, students experience difficulty in reading, analyzing, and evaluating informational text (Best et al., 2008; Cummins, 2013; Dennis, 2013; McNamara et al., 2011; Sanacore & Palumbo, 2009; Thompson et al., 2012; Zabrucky & Ratner, 1992). Researchers have identified multiple related factors that likely contribute to the gap in readers’ comprehension of literary and informational text. For instance, reading experts agree that differences in the content, structure, and complexity of informational and literary texts require readers to apply different knowledge and reading skills for successful comprehension (Duke, 2000; Duke & Kays, 1998; Duke & Roberts, 2010; Pappas, 1991). Additionally, various student characteristics such as motivation (Bulgren, Graner, & Deshler, 2013; Denton et al., 2015; Logan, Medford, & Hughes, 2011), world knowledge (Kintsch, 2013), and reading ability (Sáenz & Fuchs, 2002) can affect comprehension differently across readers.
Another reason cited for the differences in students’ comprehension of literary and informational text is the lack of exposure and use of informational text in primary classrooms. Students in the U.S. traditionally learned to read predominately with literary texts (Adams, 1990; Dewitz, Jones, & Leahy, 2009; Donaldson, 2011; Duke, 2000; Duke, Bennett-Armistead, & Roberts, 2003; Egan, 1998; Jeong et al., 2010; Kletzien & Dreher, 2004; Pressley, 2002; Stephens, 2007). In a seminal article on this point, Duke (2000) investigated the amount of informational text exposure and instruction present in first-grade classrooms and found that literary text was overwhelmingly dominant over informational text. For example, in the classroom studied, a mean of only 2.6% of walls and other surfaces, 9.8% of classroom libraries contained informational text, and 3.6 minutes per day were spent with informational texts during written language activities. Ten years later, Jeong et al. (2010) found similar results in a study investigating the availability and use of informational text in second-, third-, and fourth-grade classrooms. By the time students reach sixth grade and continuing through high school, however, 75% of texts used are non-literary (Moss, 2004). Furthermore, some analyses indicate that up to 90% of texts encountered by adults, whether in print or digital formats, are informational (Smith, 2000; White, Chen, & Forsyth, 2010). Students who experience difficulty with informational text comprehension in Grades 4-12 will more than likely face similar challenges with this text type as adults within the workplace and in their personal lives (Duke, 2000; Organisation of Economic Co-Operation and Development, 2013).

Once in the fourth grade, research indicates that an increasing number of students begin to encounter challenges with reading comprehension (Chall, Jacobs, & Baldwin, 1990; Sweet & Snow, 2003). Many reading experts have attributed this decline to the shift in instructional focus from learning how to read to using reading to learn in the content areas (Myers & Savage, 2005;
Snow, 2002). This change means readers who had success in the early grades with literary passages may struggle in the middle/secondary grades due to the increased complexity of the words, text structures, and comprehension tasks associated with informational texts (Biancarosa & Snow, 2006; Bowen, 1999; Chall & Jacobs, 2003; Duke, Halliday, & Roberts, 2013; Newkirk, 1989; Sanacore & Palumbo, 2009; Snow, 2002). Multiple studies have reported anywhere from 13% to 46% of struggling readers are first identified after the primary grades (e.g., Badian, 1999; Catts, Compton, Tomblin, & Bridges, 2012; Leach, Scarborough, & Rescorla, 2003). These young adolescents are considered to have late-emerging reading disabilities (Catts et al., 2012; Chall & Jacobs, 1983; Leach et al., 2003; Lipka, Lesaux, & Siegel, 2006; Speece, Ritchey, Silverman, Schatschneider, & Andrusik, 2010). Of particular importance, Kieffer (2010) posited that students from disadvantaged socioeconomic backgrounds are at a higher risk of developing reading problems in late elementary and middle school. In another study, Etmanskie, Partanen, and Siegel (2016) found that fourth-grade males performed worse on reading comprehension than females and that males show persistent reading difficulties across elementary and middle-level years compared to their female counterparts.

Given the contribution of informational reading to academic achievement and post-high school outcomes (Moss, 2005; Palincsar & Duke, 2004), reading curriculum and assessment developers now place increasing emphasis on this text type (Duke, 2000; Jeong et al., 2010; Kamil & Lane, 1997; Pappas, 1993). The Common Core State Standards for English Language Arts/Literacy (CCSS ELA/Literacy), for example, classify text into two broad groups. The literature category comprises stories, dramas, and poetry, whereas the informational text category includes literary nonfiction and historical, scientific, and technical texts (NGA Center & CCSSO, 2010a). CCSS ELA/Literacy documents call on schools to follow the recommendations of the
NAEP Reading Framework (NAGB, 2015), which prescribes the use of 50% informational text in Kindergarten through Grade 4, 55% in Grades 5 through 8, and 70% in Grades 9 through 12 (NGA Center & CCSSO, 2010a). Although the terms used to refer to the two text types vary, for this study, the terms literary and informational are used.

With NCLB accountability legislation, a process called Response to Intervention (RtI) was presented within the notes of regulations implementing the reauthorization of the Individuals with Disabilities Education Improvement Act (IDEIA) of 2004 (Assistance to States, 2006). The purpose of RtI is to employ a multi-tiered system of supports (MTSS) to meet the needs of struggling learners (Cortiella & Horowitz, 2014; Fuchs, Fuchs, & Compton, 2012; Jimerson, Burns, & VanDerHeyden, 2016). In response to the federal and state directives, many districts have developed MTSS frameworks (Freeman, Miller, & Newcomer, 2015) that include assessment systems and decision-making processes used to identify struggling readers and target specific areas of need for reading intervention (Brown-Chidsey & Steege, 2010; Johnson, Jenkins, Petscher, & Catts, 2009; Johnson, Mellard, Fuchs, & McKnight, 2006; Mellard & Johnson, 2008; Wayman, Wallace, Wiley, Tichá, & Espin, 2007). An intervention model uses universal screening to determine which students are at risk for not meeting academic goals (Johnson et al., 2009; Mellard et al., 2009) and predict those who are not likely to reach proficiency levels on state assessments (Hintz & Silberglitt, 2005). Based on further analysis of diagnostic assessment data, practitioners provide targeted reading intervention at increasing levels of intensity to those students who are in need. Finally, students are monitored using multiple measures to evaluate progress and inform decision making (Brown-Chidsey & Steege, 2010; Cortiella & Horowitz, 2014; Jimerson et al., 2016).
A critical factor in the success of MTSS is the use of technically adequate data sources that contribute to valid and reliable decision-making around student achievement (Busch & Reschly, 2007; Shapiro, Solari, & Petscher, 2008). First, screening tools must be sensitive to both short- and long-term changes in student performance (Haager, Klingner, & Vaughn, 2007). Second, assessments must guide practitioners when making meaningful decisions about instruction and curriculum planning (Fuchs & Fuchs, 2006). Third, assessments have to be efficient and effective (i.e., quick and easy to implement and score, cost-effective) for schools to screen all students and monitor individual student’s response to instruction (Haager et al., 2007).

Schools employ a variety of methods for universal screening, including the type and number of screening tools used (Jenkins, Schiller, Backorby, Thayer, & Tilly, 2013; Prewett et al., 2012). I discuss the two most popular classes of tools used for universal screening, general outcome measures (GOMs, e.g., curriculum-based measurement [CBM]) and computer adaptive tests (CATs), in the review of the literature (Chapter 2).

In this accountability era, with the increased use of data for high-stakes decision making, many standards-based state assessments have evolved in their application beyond their intended design (Heubert & Hauser, 1999). Historically, scores from these tests were used for accountability purposes by providing information about aggregated performance at the state, district, and school level. More recently, the results of high-stakes state assessment have been interpreted at the individual student level as well (Dennis, 2013; Lesaux & Kiefferm 2010; Rupp & Lesaux, 2006). Despite the increased expectation to use data-based decision-making, few educators are adequately prepared to efficiently and effectively navigate the abundance of information available to them (Crone et al., 2016). For example, the primary score reported in reading is a composite measure of comprehension. Some researchers have questioned the
educational utility of composite scores on these assessments for informing instructional and intervention plans (Buly & Valencia, 2002; Linn, 2000; National Research Council, 2001). The assumption is that the underlying reading problems for students who do not meet proficiency are homogeneous; therefore, most students need a similar intervention. Buly and Valencia’s (2002) research probed past failing scores on state tests and found that reading failure is multidimensional and individual. They recommended that schools should guard against superficial interpretations of and responses to state assessment results. Other experts have urged practitioners to conduct a fine-grained analysis of the problem (Dennis, 2013). That is judgments grounded in data, statistical models, and prediction models outperform those based mostly on professional intuition (Smolkowski & Cummings, 2015).

Most commercially available, standardized measures of reading comprehension include intertwined passages of literary and informational text making separate scores for each impossible (e.g., Gates MacGinitie Reading Tests–Reading Comprehension [GMRT–RC; MacGinitie, Maria, & Dyer, 2000]). Thus, few studies have directly compared student performance on standardized assessments of informational text versus literary text (Martin & Duke, 2011). Due to the lack of available standardized assessments that include separate scores for literary and informational passages, researchers have mostly used experimental assessments to investigate how learners process and comprehend texts differently (Denton et al., 2015; Kulesz, Francis, Fletcher, & Barnes, 2016).

Standardized tests may be needed that provide separate estimates of performance on literary versus informational text. Considering the increased instructional focus on informational texts (Witmer, Duke, Billman, & Betts, 2014), identifying those students who experience difficulty specifically with comprehending informational text will enable teachers to target
instruction to address individual student needs within this area. Although assessment systems measuring different components of reading have evolved within the past few decades (Ball & Christ, 2012), few technically adequate screening tools are available to identify students who experience difficulty with informational text comprehension as a separable component from literary text comprehension (Witmer et al., 2014). The Center on Response to Intervention at the American Institute for Research in collaboration with the National Center on Intensive Intervention has evaluated the technical rigor of universal screening tools. Currently, only two measures that are included in the tools chart, Northwest Evaluation Association Measures of Academic Progress (NWEA MAP, 2011) and i-Ready Diagnostic by Curriculum Associates, LLC (2015), have separate scores for each text type. Although subscores are provided, analyses to date have only investigated the predictive validity and classification accuracy of reading composite scores.

**Purpose of the Study and Research Questions**

The primary purpose of this exploratory study was to examine the diagnostic utility of the Measures of Academic Progress for Reading (MAP-R, Northwest Evaluation Association [NWEA], 2016) and demographic variables in predicting student performance on the eMPowerME English Language Arts/Literacy (eMPowerME ELA/L, Measured Progress, 2016). Additionally, this study was designed to identify the minimum MAP-R scores needed in the spring of Grade 6 and 7 that demonstrated a student was on track to meet proficiency standards on the eMPowerME one year later in Grade 7 and 8. As far as I know, there are no published studies to date that examine the predictive utility of a universal screener in predicting performance on the eMPowerME ELA/L test.
Although prior research using experimental tasks indicated that students were more likely to experience difficulty with informational more than literary text, the assessment literature has been limited regarding the performance of students on standardized measures of informational text comprehension that may be used by schools for instructional decision-making within multi-tiered frameworks. The MAP-R includes an overall score and subtest scores in Literature, Informational Text, and Vocabulary Acquisition and Use. A secondary purpose was to determine whether differences existed between middle-level students’ subtests scores on the MAP-R.

The overall aim was to inform universal screening practices at the middle-level by identifying students at risk of not meeting proficiency levels on the state assessment and, subsequently, grade-level goals for comprehending different text types. The investigation addressed the following research questions:

1. What is the concurrent validity of the MAP-R and the eMPowerME ELA/L in Grades 7 and 8?
2. How useful are MAP-R scores in identifying sixth and seventh graders who will and will not meet proficiency standards on eMPowerME ELA/L one year later?
3. Is the predictive validity of the MAP-R scores improved with the inclusion of demographic variables (i.e., gender, SES status, disability status)?
4. What are the optimal cut scores for MAP-R in spring of Grades 6 and 7 for predicting performance on the eMPowerME ELA/L in Grades 7 and 8?
5. Are there differences between and among subtest scores on the MAP-R for sixth, seventh, and eighth graders, and if so, do the differences vary according to student demographics (i.e., gender, SES status, disability status)?
6. Do MAP-R literary text and informational text scores provide unique
contributions explaining variance in eMPowerME ELA/L performance for sixth-seventh and seventh-eighth graders?

As explained in more detail in Chapter 3, I employed correlation, logistic regression, and receiver operating characteristic (ROC) curves analyses to examine the relationship between scores on the MAP-R and scores on the spring 2017 administration of eMPowerME ELA/L. In addition to MAP-R as a single predictor variable, I investigated whether results varied when student demographics were added into the model. I used a one-way repeated measures analysis of variance (ANOVA) to determine whether there were any statistically significant differences between student performance on the three MAP-R subtests and independent-samples t-tests to compare whether there were any differences among subtests according to student characteristics. Finally, I conducted a hierarchical logistic regression to ascertain the contribution of each MAP-R subtest in explaining unique variance in performance on the eMPowerME ELA/L.

**Significance of the Study**

The present investigation was designed to make contributions in the following ways. First, it aimed to contribute to the growing knowledge base in the field of adolescent reading assessment research. The results of the study have implications for researchers, policymakers, test developers, and practitioners within various educational contexts in Maine and across the country. For instance, findings might give guidance to school leaders when making decisions about the purpose and use of universal screening and benchmark assessments at the middle level. Results may add to the evidence-base regarding technically adequate decision-making within MTSS, as well as add to the validity evidence for both the MAP-R and eMPowerME ELA/L tests. Middle-level education teams could potentially use the information to inform curriculum
and instructional goals. Finally, this investigation could shed light on the importance of reviewing assessment research when considering policy decisions.

**Organization of the Study**

This study is organized into five chapters, including this chapter as an introduction to the study. In chapter 2, I review the relevant literature focusing on adolescent reading achievement in the middle grades generally and more specifically, on the reading comprehension of literary and informational text. I further discuss the common methods for universal screening that are used in schools as well as the methods and standards for classification accuracy analyses. In chapter 3, I present the design of the study, including a description of the participants, data collection, and procedures for data analysis. In chapter 4, I report the results of the analyses and Chapter 5, discuss the key findings, limitations, and implications of the study for practice and future research.
Chapter 2

REVIEW OF THE LITERATURE

Reading achievement is an essential component of college and workforce readiness (American College Test [ACT], 2004, 2006a, 2006b; Achieve, 2004; Wyatt et al., 2011), still a large number of adolescents in the United States (U.S.) struggle with comprehending the texts they read in school (Hock & Deshler, 2003; National Center for Educational Statistics, 2013). Trends in large-scale national assessment data show students’ lack of readiness to perform adequately in the world of work or college is evident well before leaving high school (Kena et al., 2015; Mattern et al., 2014; McFarland et al., 2019; U.S. Aud et al., 2013). Many students across the U.S. perform below grade level in reading when they enter the middle grades and continue to fall short in literacy achievement throughout their high school careers (Rampey, Dion, & Donahue, 2009). Not unexpectedly, concerns about the reading performance of students in grades four through twelve in recent years has directed national attention toward adolescent literacy achievement (ACT, 2008; Biancarosa & Snow, 2006; Carnegie Council on Advancing Adolescent Literacy, 2010; Center on Education Policy, 2007; Fisher & Ivey, 2006; Heller & Greenleaf, 2007; International Reading Association, 2002; McFarland et al., 2019).

Some reading experts suggested the increased expectation to comprehend informational text as a leading cause for the decline in overall reading achievement (Biancarosa & Snow, 2006; Chall & Jacobs, 2003; Chall, Jacobs, & Baldwin, 1990; Duke, Halliday, & Roberts, 2013; Newkirk, 1989; Sanacore & Palumbo, 2009). Seen around the fourth grade, many students experience difficulty with extracting meaning from informational text (Chall et al., 1990). Because researchers have shown that students generally struggle with comprehending informational text over literary text (Best et al., 2008; McNamara et al., 2011; Thompson et al.,
and due to the increased access to informational text in the twenty-first century, the need to improve informational text comprehension is now gaining widespread attention (NGA Center & CCSSO, 2010).

In the last two decades, the greater emphasis on accountability has led districts to develop Response to Intervention (RtI) frameworks (Freeman et al., 2015) that include assessment systems and decision-making processes used to identify struggling learners and target specific areas of need for intervention (Brown-Chidsey & Steege, 2010; Johnson et al., 2009; Johnson et al., 2006; Mellard & Johnson, 2008; Wayman et al., 2007). Considered as a general education initiative, RtI calls for large-scale improvements in academic and behavior outcomes for all students.

A broad purpose of RtI is to merely act as a multi-level system directed toward supporting students who are struggling (Fuchs et al., 2012). Beneath the surface, though, RtI is a complex process for implementing high-quality, research-based interventions based on learner needs. Major component of RtI include universal screening of all students for academics and behavior, monitoring progress of student performance, adjusting teacher instruction based on student response, checking the fidelity of implementation, and making team-based educational decisions that are data-driven and can be applied to general, remedial, and special education (Brown-Chidsey & Steege, 2010; Jimerson et al., 2007; Johnson et al., 2006, Mellard & Johnson, 2008). The goal is to have a seamless assessment and instructional delivery system based on students’ response.

A major outcome of properly instituted RtI methods is early identification and intervention for students who struggle academically and behaviorally (Bradley, Danielson, & Doolittle, 2007; Brown-Chidsey & Steege, 2010; Fuchs & Fuchs, 2006, 2007; Fuchs & Deshler,
The service delivery components of RtI are based on thirty years of research in behavioral consultation (e.g., Bergen, 1977), data-based problem identification (e.g., Deno, 1985; Deno & Mirkin, 1977), general-outcome measures ([GOMs] e.g., curriculum-based measurement [CBM] Deno, 1985, Shinn, 2008), evidence-based practices (e.g., Vaughn, Linan-Thompson, & Hickman, 2003), multi-tiered systems of support ([MTSS] e.g., Walker & Shinn, 2010), and functional behavior assessment and analysis (e.g., Gresham, 2005).

In this chapter, I review the research on adolescent reading achievement in the middle grades generally and more specifically, on the reading comprehension of literary and informational text. First, I provide a brief overview of the historic and current focus on informational text. Then, I examine the differences between informational and literary text on characteristics such as content, structures, features, and cohesiveness, and describes reports and studies that directly compared student performance on informational versus literary text comprehension. This review also includes an analysis of student performance on large-scale assessments as well as reader and text characteristic associated with text comprehension and summarizes the factors affecting informational text comprehension. Finally, I close the chapter by reviewing the current literature on the technical adequacy of reading universal screeners in identify students at risk for reading failure and predicting performance on high-stakes state assessments.

**Historical and Current Focus on Informational Text**

Historically, students in the U.S. learn to read predominately with literary texts (Dewitz et al., 2009; Donaldson, 2011; Kletzien & Dreher, 2004; Stephens, 2007) and are not adequately prepared to engage with informational text in the upper elementary grades and beyond (Duke,
The shift in focus toward reading for understanding generally happens after the initial stages of reading acquisition in the third grade. At this time, instruction in the content areas increases the exposure to informational text (Bowen, 1999; Snow, 2002). By the time students reach sixth grade, and beyond, 75% of texts used are non-literary (Moss, 2004). Some figures show that up to 90% of texts encountered by adults are informational (Smith, 2000).

In recent decades, some reading experts realized the importance of focusing instruction on informational text in the lower grades to prepare students for increasing the complexity of this text type in the upper grades (Caswell & Duke, 1998; Duke, 2004; Hirsch, 2003). The authors of the CCSS claim the standards will prepare students college and career ready by improving comprehension of informational text across the K-12 curriculum (NGA, 2010b). Following the 2009 Reading Framework, the standards placed greater emphasis on informational text at the secondary level (NAGB, 2015). Critics of the CCSS suggested students would place less value on literature and questioned whether or not informational reading in high school promoted college readiness (Bauerlein & Stotsky, 2012).

**Informational and Literary Text Characteristics**

Reading experts agree that differences in the content and structure of informational texts and literary texts require readers to apply different knowledge and reading skills for successful comprehension (Duke, 2000; Duke & Kays, 1998; Duke & Roberts, 2010; Pappas, 1991). The qualitative dimensions of text complexity, as described in Appendix A of the CCSS ELA/Literacy, show dissimilarities between the text types in levels of meaning or purpose, structure, language conventionality and clarity, and knowledge demands (NGA Center & CCSSO, 2010b). Literary texts typically follow a single structural pattern often referred to as
story grammar or rules devised for expressing the structure of stories (Mandler & Johnson, 1977). These passages usually follow a linear progression of related events, have predictable story elements of setting, character, and plot, are written in the past tense, and use familiar vocabulary. Furthermore, literary texts usually present recurring themes (e.g., love, friendships, coming of age, person versus nature). Hence, readers often have extensive experience and knowledge (i.e., schemas) about the settings, actions, and events described by literary texts (Nelson, 1996; Olson, 1985). For that reason, some researchers assert that most children possess adequate event-related knowledge to comprehend literary texts successfully (Best, Ozuru, Floyd, & McNamara, 2006; Best et al., 2008).

Informational texts, on the other hand, comprise a variety of structures that are more complex and variable. Although the terms may differ in the literature (Anderson & Armbruster, 1984; Meyer, Brandt, & Bluth, 1980), the structures of informational text are generally referred to as (a) description, (b) sequence, (c) cause-effect, (d) compare-contrast, and (e) problem-solution (Reutzel, Jones, Clark, & Kumar, 2016). These structures may be applied solely or in combination within a single informational text. Findings from past and present research studies and curriculum documents have shown that explicit teaching of text structures and features of informational texts improves students’ comprehension (Duke, Pearson, Strachan, & Billman, 2011; Pearson & Fielding, 1991; Shanahan, 2011; Williams, 2005).

In addition to structural differences between literary and informational text, there are structural variations across content areas within informational text. Different disciplines represent unique abstract concepts and vocabulary, which can be difficult to interpret for many students (Hynd, 1998). For example, Armbruster and Nagy (1992) determined that unknown words encountered in informational texts are more conceptually challenging, likely more
interrelated thematically, and ultimately interfere more with meaning-making than in literary texts. Similarly, in another study, Hiebert and Cervetti (2011) examined the words selected for instruction from fourth-grade English/language arts and science programs and found the words in the literary vocabulary are more likely to be familiar to students than science words. The science words were significantly longer and had definitions conceptually more complex than literary words. In general, researchers have found strong correlations between vocabulary knowledge and reading ability in childhood (Quinn, Wagner, Petscher, & Lopez, 2015; Snow, Porche, Tabors, & Harris, 2007; Tannenbaum, Torgesen, & Wagner, 2006) and even more so in adolescence (Cromley & Azevedo, 2007). Consequently, the vocabulary differences between the two text types may influence student comprehension with informational texts containing more complex, conceptual vocabulary than literary texts.

The coherence of a text is another key factor in comprehension (McNamara, Kintsch, Songer, & Kintsch, 1996). Text cohesion relates to the degree in which the text provides background information and explicit cues that direct the reader to connect information presented in different parts of the text (Graesser, McNamara, & Louwerse, 2003). Whereas high-cohesion texts contain multiple clues to relations within and across sentences, low-cohesion texts have minimal clues and require readers to make many inferences to construct a coherent representation of the text (McNamara, Louwerse, McCarthy, & Graesser, 2010). Although both literary and informational texts can exhibit varying levels of cohesiveness, low-cohesion informational passages are generally considered to be more knowledge demanding for readers than high-cohesion literary passages (Beck, McKeown, & Gromoll, 1989). For example, Beck et al. (1989) completed a comprehensive analysis of four elementary social studies texts and found
the texts had unclear goals, minimal links, and assumed too much background knowledge on the part of readers.

Interestingly, studies have demonstrated that after revisions were made to improve the structural and explanatory coherence of passages, readers’ recall of the text increased (Beck, McKeown, Sinatra, & Loxterman, 1991; Britton & Gulgoz, 1991). Kintsch (1990), on the other hand, found that high-cohesion texts reduce the amount of active processing during reading. That is, after the demands for figuring out the meaning of the text were eased, less learning occurred for knowledgeable and skilled readers. Kintsch determined the quality of learning from a text by analyzing the students’ written summaries and concluded that the students wrote better summaries of a poorly organized text than a well-organized one. Kintsch’s (1988, 1992, 1998, 2004) Construction-Integration (CI) model of text comprehension and similar process models underscore the somewhat contradictory findings regarding easy to read, high-cohesion texts being generally easier to recall while low-cohesion, difficult-to-read texts require active processing and thus facilitates learning. Proponents of process models of reading comprehension like the CI model argue that successful reading comprehension requires the construction of a coherent mental representation of the text being processed (Graesser, Singer, & Trabasso, 1994; Yeari & van den Broek, 2011).

The CI model distinguishes between different levels of processing in the mental representation of text that readers construct. The surface code consists of the exact words and syntax of clauses and typically quickly fades from memory. The phase of processing assumes the reader can accurately decode and recognize words. The textbase involves the reader translating the words into meaning. At the deepest level of text processing, the situation model moves beyond the explicit meaning of the text content. The reader integrates the meaning of the text and
world knowledge to gain the global meaning of the text. Readers typically form a similar
textbase; however, situation models vary across readers and text types. The situation model for
literary texts includes the reader’s understanding of characters, settings, actions, and events. For
informational text, the situation model refers to the integration of the textbase and the reader’s
knowledge about the subject matter (Graesser et al., 1994; Kintsch, 2004).

In summary, researchers have demonstrated that the variance in informational text
features including structure, content, vocabulary, and cohesiveness make it typically more
challenging for most students to comprehend than literary texts (Best et al., 2008). The role text
cohesion plays in reading comprehension is discussed further in the next section of this review.

Comparing the Comprehension of Informational Versus Literary Text

Research evidence suggests that students generally have more difficulty comprehending
informational text than literary text (Best et al., 2008; McNamara et al., 2011; Thompson et al.,
2012). The two largest data sources that enable comparison of student performance on
informational versus literary passages are the Progress in International Reading Literacy Study
(PIRLS) and the National Assessment of Educational Progress (NAEP). An international
comparative assessment, including 54 education systems that evaluates the performance of
fourth-graders in reading, the PIRLS has been conducted every five years since 2001 (Mullis &
Martin, 2015). PIRLS yields conclusions about student performance on informational versus
literary passages by comparing performance on each score to student’s overall composite score
in reading. According to the 2011 PIRLS data, the mean literary text scale score was
significantly higher, and the mean informational text scale score was significantly lower than the
average composite reading score of U. S. fourth graders (Mullis, Martin, Foy, & Drucker, 2012).
NAEP has assessed reading achievement in representative samples of U.S. students in Grades 4, 8, and 12 since 1992 (NCES, 2013). NCES publishes reports summarizing key trends on NAEP and also allows users to explore additional questions regarding student achievement using the NAEP Data Explorer. This tool is a free, web-based system that generates tables and detailed results of interest (NCES, 2019). Although NAEP generates two subscores, one for informational and one for literary text, results can only be used to analyze trends for each text type separately (e.g., comparing within-text type differences by gender, grade, race, or socioeconomic status [SES]). For instance, fourth and eighth-grade students who were not eligible for the national school lunch program significantly outperformed students who were eligible (p < .05) on both text types according to the 2017 results (NCES, 2019).

In another example, Klecker (2006, 2014) conducted secondary analyses on NAEP testing and found females in fourth and eighth grade significantly outscored males on both text types for every NAEP testing year except 1998. Furthermore, the mean difference between scores for girls and boys at both grade levels was consistently higher for literary text than for informational. These NAEP data are congruent with PIRLS wherein fourth-grade girls consistently performed higher on literary reading tasks over fourth-grade boys, and the achievement differences were smaller between genders for informational reading but still significant (Mullis et al., 2012).

Large-scale assessments shed some light on how students generally comprehend informational and literary text differently. Unfortunately, NAEP reports do not directly compare performance on the two-component scores for the overall population nor does NAEP data explorer allow users to generate statistical analyses that directly compare performance on the two subscores. Because neither the NAEP nor international studies make direct comparisons,
researchers cannot draw any conclusions on whether U.S. students comprehend one kind of text better than the other based on the results of these assessments.

Multiple factors may contribute to the difficulty students experience with comprehending informational and literary texts. Component reading skills such as decoding and word reading skills, prior knowledge, and vocabulary knowledge (Paris, 2005) may impact comprehension differently according to text type. Word reading and reading comprehension are strongly correlated (Ouellette, 2006). Indeed, students with low word reading skills tend to experience difficulty with making sense out of what they read (Dennis, 2013). Prior knowledge or world knowledge is another essential factor that contributes to reading comprehension. Higher knowledge readers comprehend texts better than readers with little prior knowledge (McNamara et al., 2011). For instance, Best and colleagues (2008) examined the influences of reading decoding skills and world knowledge on typically developing third graders’ comprehension of literary and informational text. Text passages were analyzed and equated via quantitative measures, including the number of words, the number of sentences, Flesch-Kincaid grade level, and Flesch reading ease. After reading one passage of each type, student comprehension was assessed employing a free recall prompt, three cued recall prompts, and 12 multiple-choice questions. Findings revealed that comprehension was better for the literary than the informational text. Decoding skills most influenced literary comprehension, whereas world knowledge most influenced comprehension of informational passages.

In another study, McNamara and colleagues (2011) examined the roles of reading decoding skills and world knowledge among high- and low-knowledge fourth grade students after reading both high- and low-cohesion literary and informational passages. The overarching purpose of the study was to contribute to the existing literature base regarding the emergence of
reading comprehension difficulties that typically occur for students around the fourth grade. Comprehension was assessed using free and cued recall and multiple-choice questions. Results revealed that informational text was more difficult than literary text, and high-knowledge readers comprehend better than low-knowledge readers across both text and cohesion types. Effects for knowledge were higher for informational text for both low- and high-knowledge students.

Some studies found in the literature investigated the effects of cognitive ability on comprehension according to the genre and found that informational texts were more difficult for readers regardless of their overall ability and reading achievement levels. Research has established that student with identified learning disabilities experiences greater difficulty engaging with and making meaning from informational text (e.g., Englert & Hiebert, 1984; Wong & Wilson, 1984). For example, Sáenz and Fuchs (2002) found that high school students (Grades 9-12) identified with a learning disability (LD) read not only informational text less fluently than narrative text but also comprehend informational text less well. Other researchers have investigated the comprehension of students without identified learning disabilities but with low reading achievement. For instance, Olsen (1985) explored third-graders' ability to answer text-based inference and paraphrase questions after reading literary and informational passages. Participants of the study were classified as good or poor readers based on the results of a recent battery of achievement testing. More specifically, the good readers scored average or above on all measures while the poor readers scored average or above on all measures except for scoring below average on the comprehension subtest. Results indicated that informational passages were significantly more challenging to understand than literary for both good and poor readers.

Using different methodologies, other researchers conducted similar studies with middle and secondary students (Berkowitz & Taylor, 1981; Denton et al., 2015; Zabrucky & Ratner,
1992). In all cases, results were similar showing that both good and poor adolescent readers were able to recall significantly more information from literary passages than from informational. Of relevance to the proposed study, continuing research is needed to examine the text processing difficulties that characterize a large percentage of adolescent readers. Valid decision-making based on assessment results is needed in order to identify those students who struggle with informational text in particular. Understanding how passage characteristics interact with student characteristics is essential for the development of effective instruction.

A small body of research has indicated that text difficulty or the way passages are written could affect readers’ ability to make meaning from different types of text (Barth, Tolar, Fletcher, & Francis, 2014; McNamara, 2001; McNamara et al., 1996; McNamara & Kintsch, 1996). Specific findings in the McNamara et al. (2011) study related to text cohesion differences showed that students comprehended high-cohesion literary texts better than low-cohesion texts, whereas there was no effect of cohesion for the informational passages. Further analysis revealed that high-knowledge students tended to understand the low-cohesion literary text better than high-cohesion literary texts indicating a reverse cohesion effect.

In a similar study, Denton et al. (2015) implemented a think-aloud methodology to examine students’ conscious engagement in inference generation, paraphrasing, verbatim text repetition, and monitoring while reading texts that were either at or above the students’ reading level. Informational and literary passages were selected from various literary works, philosophical and science texts, and political essays. The readability of each passage was determined by using Lexile levels and then matched to the readers’ Lexile level (Schnick, & Knickelbine, 2007). They determined that text difficulty did not have significantly stronger effects for poor comprehenders relative to adequate comprehenders. The processing of poor
comprehenders was impacted primarily by genre with informational text posing significant more
difficulty for them than literary passages. Subsequent examination of adequate comprehenders
revealed they generated more inferences when reading accessible literary passages than when
they read challenging literary text, whereas inference rates were similar when reading both
accessible and challenging informational text.

In the studies mentioned above that investigated the effects of text difficulty on learners’
comprehension, many conclusions can be drawn (Denton et al., 2015; McNamara et al., 2011).
First, comprehension problems become more evident in the fourth graders who have not gained
sufficient knowledge about the world. Second, text selection matters regardless of reading
comprehension ability, decoding skills, or world knowledge of adolescent readers. Third,
adolescent readers may benefit from instruction in inference generation and paraphrasing that
begins with high-cohesion literary texts and progresses systematically to low-cohesion literary
text and informational text.

The ability to make inferences is essential to comprehension. Readers are required to fill
the gaps when information is not explicitly stated in a text. Different questions draw upon
various comprehension processes (Eason et al., 2012). Because a text is not always explicit, the
properties of the questions being asked is another consideration used by some researchers when
comparing readers’ comprehension of literary and informational passages. Researchers have
described various inference taxonomies in the literature that parse out the different functions
inference plays in comprehension (Nicholas & Trabasso, 1980; Pearson & Johnson, 1978;
Warren, Nicholas, & Trabasso, 1989). Simply put, literal questions assess readers’ ability to
recall information explicitly stated in the text (Kintsch, 1994). Inferential questions or schema-
based inferences, on the other hand, require readers to either (1) piece together related
information that is presented in multiple areas in the text (i.e., text-based inferences) or (2) combine prior knowledge with the information from the passage (i.e., knowledge-based inferences; Hannon, 2012).

In the Olson (1985) study, students were asked questions of varying difficulty (i.e., paraphrase, informational inference, logical inference) with paraphrase considered as the easiest and logical inference as to the most challenging. Results indicated that overall, informational passages were significantly more difficult to understand than literary passages for both groups of students across all question types, and good readers answered significantly more questions than poor readers on both types of text. Furthermore, paraphrase and informational inference questions were equally difficult for both good and poor readers after reading literary and informational passages. Interestingly, students found logical inference questions significantly more difficult than informational inference questions only after reading literary passages. Put another way; logical inferences were not significantly more complicated than the other categories of questions for both good and poor third-grade readers after reading informational text.

Comparable to Olson’s (1985) findings, Saenz and Fuchs (2002) found in their investigation that secondary students with LD had similar literal comprehension on informational and literary passages. A somewhat different result than Olson’s, however, LD students had poorer inferential comprehension on informational text (Cohen’s $d = 0.42$). One plausible explanation for this disparity in findings could be that Olson (1985) examined informational and logical inference types questions separately, whereas Saenz and Fuchs (2002) analyzed the two types of inference questions as a single category.

In another study, Kulesz and colleagues (2016) used passages from the GMRT-TC (MacGinitie et al., 2000) to investigate the comprehensive and systematic effects of passage
characteristics, question types, and component reading skills on secondary students' reading comprehension. Because the test publisher did not differentiate the literary and informational passages, however, the researchers analyzed and coded the passages according to text type, overall passage difficulty, and text cohesion as measured by the Lexile score of the passage (Schnick, & Knickelbine, 2007) and Coh-Metrix Text Easability Assessor (Graesser, McNamara, & Kulikowich, 2011), respectively. The participants in this sample were coded as high or low comprehenders with adequate decoding skills as determined by the state reading comprehension test and a standardized letter-word identification measure. From their analysis, they concluded that informational passages were more complicated than literary passages across all participants. In contrast with findings of the two studies mentioned above (Olson, 1985; Saenz & Fuchs, 2002), the results suggested that literal questions on average were more difficult than text-based inferences. The authors determined that this counterintuitive finding stemmed from the fact that passages in which literal questions were drawn had less text cohesion than those passages with questions requiring text-based inference.

**Summary of Informational Text**

The ability to read, analyze, and evaluate informational text is essential for college and the working world (Cummins, 2013; Duke et al., 2013). As such, prior research has indicated that as a heterogeneous group, most students tend to fare better on measures of comprehension of literary passages than on informational texts (e.g., Best et al., 2008; Dennis, 2013; McNamara et al., 2011; Thompson et al., 2012). Multiple factors such as learner characteristics, text characteristics, and assessment approaches contribute to the difficulty with comprehending informational passages. For instance, researchers have shown the impact world knowledge has on comprehension of informational text (Beck & McKeown, 1992; Freebody & Anderson, 1983;
Johnston, 1984; Kintsch, 2013; Langer, 1986; Lipson, 1982; Yochum, 1991). Other experts have identified decoding skills (Lyon, 2002; Vellutino, 2003), vocabulary knowledge (Hiebert & Cervetti, 2011; Sanacore & Palumbo, 2009), and motivation (Bulgren et al., 2013; Denton et al., 2015) as predictive factors. Other notable factors cited in the literature that may affect comprehension of informational passages include the level of text complexity or text cohesion (Duke et al., 2011; McNamara et al., 2011) and the level of exposure to different types of text during comprehension instruction of literary and informational text (Duke, 2000). The level of difficulty of the comprehension tasks and question types are also variables that influence the comprehension of informational text (Eason et al., 2012). Students' academic achievement and successes in life are much dependent upon the ability to understand a wide range of texts (Chall, 1996; Snow, 2002). Not only is reading comprehension a critical foundation for learning and knowledge, but also it is an essential component of college and workforce readiness (Achieve, 2004; ACT, 2004, 2006a, 2006b; Wyatt et al., 2011).

**Technical Adequacy of Screening Assessments, Procedures, and Decision Making**

Although a considerable body of evidence exists on measurement approaches used to screen for academic or behavior difficulties and monitor student progress (Glover & DiPerna, 2007), additional investigations and developments are needed for ensuring that assessment and decision-making are accurate and useful. Even the most practical universal screeners are of little use without strong psychometric properties. Universal screening measures should result in data that are valid for identifying students who are at risk for later reading difficulties based on their current performance Kettler, Glover, Albers, & Feeney-Kettler, 2014). Often the evidence to support the use of universal screeners comes from correlational relationships between screening measures and more comprehensive measures of the target construct. Correlational evidence
suggests that a universal screening measure captures the construct of interest, but technical adequacy is needed to determine the usefulness of a screener (Jenkins, Hudson, & Johnson, 2007; Kilgus, Methe, Maggin, & Tomasul, 2014). A key factor in the success of RtI models is the use of technically adequate data sources that contribute to sound decision making around student achievement (Busch & Reschly, 2007; Shapiro, Solari, & Petscher, 2008). As such, educators must appropriately select valid and reliable assessment tools and implement them responsibly.

The RtI model requires that utilized assessments meet specific criteria as cited in the literature (Yeo, 2011). First, the tools must be sensitive to both short-term and long-term change in student performance (Hagaar et al., 2007). Second, assessments must guide practitioners to make meaningful decisions about instruction and curriculum planning; thus, are formative (Fuchs & Fuchs, 2006). Third, assessments have to be efficient and effective (e.g., quick and easy to implement and score, cost-effective) for schools to screen all students and progress monitor individual student response to instruction (Hagaar et al., 2007). Appropriately used, sound assessments can result in informed decision-making about individuals and programs, increase educational standards, and provide more equitable access to education (American Educational Research Association [AERA], 2014; Office of Civil Rights [OCR], 2000). There is a risk, however, that not all tests are well constructed nor are all tests used appropriately, which can lead toward harm to test takers and others affected by assessment-based decisions (AERA, 2014; OCR, 2000).

The Standards for Educational and Psychological Testing indicate three essential components that contribute to the technical adequacy of high-quality assessments as (a) validity, (b) reliability, and (c) freedom from bias (AERA, 2014). Validity is defined as "the degree to
which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9). To put it more simply, Sireci (2007) summarized four broad conclusions regarding the fundamental aspects of validity, as reported in the literature:

(a) Validity is not a property of a test. Instead, it refers to the use of a test for a particular purpose, (b) To evaluate the utility and appropriateness of a test for a particular purpose requires multiple sources of evidence, (c) If the use of a test is to be defensible for a particular purpose, sufficient evidence must be put forward to defend the use of the test for that purpose, and (d) Evaluating test validity is not a static, one-time event; it is a continuous process (p. 477).

From a RtI perspective, the validity of a reading measure, for example, is determined by the extent to which the measure serves as an indicator of general reading proficiency (Deno, 1985). In addition to the traditional conceptualization of validation (i.e., construct, content, predictive, concurrent validity), the contemporary conceptions of validity—the validity of decisions that are based on assessment results, also must be addressed (Ball & Christ, 2012). Reliability, on the other hand, refers to the degree of consistency of test scores over administrations, forms, items, scorers, and other aspects of testing (AERA, 2014). Evaluating reliability includes identifying major sources of error and the size of error resulting from these sources (OCR, 2000). Freedom from bias ensures that a test does not favor one group of test-takers over another. In other words, tests are used in a manner that "supports sound educational decisions, regardless of the race, national origin (including limited English proficiency), sex, or disability of the students affected" (OCR, 2000, p. 14).

Within the context of a RtI framework, universal screening is a necessary first step in identifying students who are at risk for academic difficulties (Johnson et al., 2009). According to
the Center on Response to Intervention (American Institutes for Research, 2007), universal screening is defined as follows, "Screening involves brief assessments that are valid, reliable, and evidence-based. They are conducted with all students or targeted groups of students to identify students who are at risk of academic failure and, therefore, likely to need additional or alternative forms of instruction to supplement the conventional general education approach."

Thus, screens are designed to predict a future outcome based on a specific criterion measure that enables schools to intervene as early as possible when necessary (Johnson et al., 2009). With that said, universal screening data can be interpreted in three ways, either norm-referenced, criterion-referenced, or self-referenced (Ardoin & Christ, 2008). First, norms allow schools to compare individuals or groups of students according to a sample of their peers (e.g., national, local). Benchmarks used in a criterion-referenced framework can help to determine how a student might perform on a large-scale test. Finally, comparing a student's score over time is useful for determining whether or not the student is responding to instruction. All three interpretations used in conjunction with one another can assist practitioners when making important decisions about student academic performance (Ardoin & Christ, 2008). First and foremost, in order for RtI to work efficiently and effectively all students that indeed are at risk must be identified and provided with the necessary supports while the number of students who are falsely identified, those students who are actually not at risk for academic problems, is limited (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996).

The accuracy of assessment measures depends upon a variety of factors such as, how well the test separates individuals into the categories of at-risk for reading problems and not at-risk for reading and whether the reliability of the performance-level score is accurate and consistent. Hence, the diagnostic accuracy refers to the ability of an instrument to distinguish
between two diagnostic alternatives and correctly select the correct one (Swets, Dawes, & Monahan, 2000). As such, effective screening tools demonstrate a high level of sensitivity in correctly identifying those students who will encounter reading difficulties, balanced with high levels of specificity in the accurate identification of those who are not likely to encounter reading problems (Zhou, Obuchowski, & Obuchowski, 2002). The ultimate goal is to maximize classification accuracy (CA). CA is the overall proportion of students who were correctly identified as at-risk (i.e., true positives) or not at-risk (i.e., true negatives) on a screening measure. CA, however, must be evaluated by considering the base rate of risk or the percentage of students identified as at-risk according to the outcome measure. Even universal screeners with excellent classification qualities may not be useful for samples with a low base rate of risk (Johnson, Jenkins, & Petscher, 2010). For instance, with a base rate of risk at 10%, declaring that all students are not at-risk without using a screener at all would result in a 90% CA (100% specificity and 0% sensitivity). To the contrary and for the same reasons, some researchers also questioned the need for a universal screener if the base rate of a sample is exceptionally high (Johnson et al., 2010).

In addition to population indices mentioned above, Christ and Nelson (2014) emphasized the importance of sample-based indices of technical adequacies, such as positive predictive value (PPV) and negative predictive value (NPV). PPV is the proportion of students identified as at-risk on the screener who ultimately fails the outcome measure, and NPV is the proportion of students who were identified not at-risk on the screener and subsequently passed the outcome measure. The actual number of students who are identified as at-risk on the outcome measure, called the base rate, can affect the PPV and NPV. Researchers and practitioners must balance the diagnostic accuracy statistics of a screening tool with the needs and goals of the district. For
example, if the aim is to efficiently and quickly identify struggling readers and provide immediate intervention, then a high level of sensitivity may be the priority. On the other hand, if the intent is to rule out those students not at-risk of reading problems as a way of excluding them from further screening, then NPV may take precedence.

Sensitivity and specificity levels can be difficult to interpret on their own and of limited practical utility without reporting associated cut scores. Receiver operating characteristic (ROC) curves is another approach for statistically evaluating the diagnostic accuracy of a universal screening measure (Smolkowski & Cummings, 2015) and determining related cut scores. A visual representation of the ROC curve, which entails plotting the true-positive rate (sensitivity) on the y-axis and the false-positive rate (1-specificity) on the x-axis for each possible screening cut score (Hilbe, 2009). The display results in a curve that begins in the lower-left corner where both proportions are 0 and rise toward the upper right corner where both are 1. A steep curve to the upper left signifies a more accurate predictor measure (i.e., higher sensitivity and higher specificity), whereas curves close to the diagonal line indicating 50% probability classification have poor diagnostic accuracy (Swet et al., 2000). The area under the curve (AUC) is an overall measure of the accuracy of the predictor. AUC is a probability index that can range from 0.5 (no diagnostic ability) to 1.0 (perfect diagnostic ability). According to Hosmer, Lemeshow, and Sturdivant (2013), an AUC value between 0.5 and 0.7 has poor discrimination, between .70 and .80 is considered acceptable, between .80 and .90 is excellent, and above .90 is regarded as outstanding discrimination.

Number of Screeners and Routes to Intervention

Schools commonly employ only a single measure within a universal screening process (Jenkins et al., 2013; Prewett et al., 2012). According to the literature, risk classification by a
single screener method has generally failed to reach an acceptable balance between sensitivity and specificity to be used for high-stakes decisions (Hintze & Silberglitt, 2005; Johnson et al., 2009). For example, Kent, Wanzek, & Yun (2018), studied the diagnostic accuracy of five separate screeners measuring word-level and text-level reading of fourth-graders and determined that all individual measures were significant predictors of student outcomes on state reading tests. Despite optimal sensitivity, however, other indices (i.e., SPE, PPV, NPV, CA) were indicative of generally inadequate technical adequacy (ranging from .15 to .88).

As an alternative to single-screener approach, researchers have investigated the use of multiple screeners to improve the technical adequacy of decision making (Fuchs & Vaughn, 2010). Multivariate screening measures combine information from multiple measures collected at the same time. Through logistic or multiple regression analysis, the screening measures are entered into the model to predict proficiency on the outcome measure. ROC curve analysis is then employed to determine a cut-score with optimal sensitivity and specificity (Catts, Fey, Zhang, & Tomblin, 2001). Several researchers reported results that show screening batteries with measures of fluency, vocabulary, and comprehension outperformed single screening measures (Baker et al., 2015; Decker et al., 2014; Shapiro et al., 2008; Speece et al., 2010). Speece and colleagues (2010), on the other hand, reported the diminishing levels of effectiveness with batteries using more than three measures.

Once students are identified as at-risk via the selected screening measure, schools are typically choosing to employ one out of two different routes to intervention (Fuchs et al., 2012). In the more commonly used approach, the direct route (DR) (Mellard, Byrd, Johnson, Tollefson, & Boesche, 2004), students are immediately placed into Tier 2 intervention based on a one-time screening measure (Fuchs & Fuchs, 2007; Fuchs, Fuchs et al., 2012; Johnson, Jenkins et al.,
This system of identification assesses all students at the beginning of the school year and targets those students who perform below a predetermined norm-referenced cut point or performance benchmark (Fuchs & Fuchs, 2007). Recent research suggests that this approach to identifying students at risk for academic problems leads to a large number of false-positives (Fuchs et al., 2012). Allocating resources for students who are indeed not at risk is an inefficient use of school funds as it compromises the efforts of practitioners trying to meet the needs of true positives (Fuchs et al., 2012).

For the other less conventional route, two-stage screening referred to as a progress monitoring route or gated-screening approach, students are first identified as potentially at-risk by way of a universal screening measure, then further progress monitored for several weeks before any Tier 2 placement decisions are made (Fuchs et al., 2012). For example, Compton, Fuchs, Fuchs, and Bryant (2006) improved the overall CA from 81% to 83% and were able to identify 90% of first graders at-risk (sensitivity) when five weeks of progress monitoring data were added to a composite screening measure. Recent studies found in the literature examined the gated-screening approach with older students. Klingbeil, Nelson, Van Norman, and Birr (2017) examined the diagnostic accuracy and efficiency of a gated-screening approach at the intermediate-level (Grades 3-5). Results showed a high level of specificity by reducing the false positive rate, thus lowering the sensitivity and subsequently increasing the number of false negatives. In a similar study, Van Norman, Nelson, and Klingbeil (2017) also found increased specificity and decreases in sensitivity when analyzing the reading scores of fourth through seventh graders.
There are pros and cons associated with each approach. For instance, the use of single-screeners and employing DR leads to expedient intervention, but may falsely identify students as at-risk or not at risk and, therefore, ineffectively distribute resources (Fuchs & Fuchs, 2007; Fuchs et al., 2012; Johnson et al., 2010). On the contrary, the two-stage model and gated approach may prolong necessary targeted or intensive intervention, but has better identification accuracy when employed correctly (Compton et al., 2006; Fuchs & Fuchs, 2007) and is less costly in the long run (Fuchs et al., 2012). More research is needed to determine what is the best approach for determining when a child should receive intervention and at which tier. For instance, some students are held in tier 2 for too long when immediate intensive or tier 3 interventions are needed (Vaughn, Denton, & Fletcher, 2010). Research studies on fast-tracking the weakest students to tier 2 or tier 3, depending on their academic skills profile are needed (Al Otaiba et al., 2014). In summary, the accuracy of universal screeners has significant implications for school districts to allocate the increasingly limited resources to those students who are at most risk for reading problems and, ultimately, at-risk of school failure.

**General Outcome Measures (GOMs)**

GOMs are standardized, brief assessments that are individually administered and used to test broad academic areas such as reading and mathematics. Educators can compare student scores against national grade-level norms for the expected level of performance and rate of improvement. Mellard et al. (2009) surveyed 41 schools and learned that 76% of them used some GOM for universal screening. A large body of research supports the utility of GOMs for predicting reading outcomes (e.g., Ardoin & Christ, 2008; Goffreda & DiPerna, 2010; Wayman et al., 2007). Acadience (formally known as DIBELS, [https://acadiencelearning.org](https://acadiencelearning.org)) and AIMSWeb ([https://www.aimsweb.com](https://www.aimsweb.com)) are examples of GOMs.
Due to its standardized procedures (Christ, Johnson-Gros & Hintze, 2005; Francis et al., 2008; Hagaar, et al., 2007), the most popular universal screening (Ball & Christ, 2012) and progress-monitoring assessment type utilized within the context of RtI for reading is curriculum-based measurement ([CBM-R] Busch & Reschly, 2007). There is substantial evidence documented in the literature indicating that CBM-R as a general outcome measure (GOM), is valid and reliable and hence, has gained acceptance as a way to screen students who may be at risk for reading failure. Leading researchers and advocates associated with CBM-R support the use of these assessments to measure general reading ability (Deno, 1985; Fuchs, Fuchs, Hosp, & Jenkins, 2001). Three comprehensive reviews of the literature summarized this documented evidence (Marston, 1989; Reschly, Busch, Betts, Deno, & Long, 2009; Wayman et al., 2007). Marston (1989) conducted the first narrative review of the literature in response to concerns related to technical adequacy and practicality of CBM reading measures. At that time, CBM was primarily used as a progress-monitoring tool to assess basic skills for students with disabilities in the elementary grades (Wayman et al., 2007). The results showed support for the use of word identification and oral reading fluency as indicators of general reading proficiency with reliability coefficients ranging mostly above .90 and validity coefficients around .80 (Marston, 1989).

In another review, Wayman and colleagues (2007) qualitatively synthesized 90 studies of CBM reading measures (CBM-R and CBM-Maze) published since Marston's (1989) seminal review, which focused on issues of technical adequacy as they relate to measures, materials, and growth. The results further supported the empirical, theoretical (Fuchs et al., 2001), and psychometric evidence for CBM-R. Reading aloud measures continued to show a strong relationship to overall reading proficiency as compared to earlier research (Marston, 1989;
Schellings, Aarnoutse, & van Leeuwe, 2006) and was found to be a better indicator of comprehension than other typical measures. Across all the studies, correlations between scores on CBM reading measures and state reading tests ranged from .60 to .80, although correlations tended to decrease at the intermediate grades. This decline in correlations in the upper grades suggests that CBM-R may not be the best measure to be used with older students. Instead, a multivariate approach that uses multiple measures and other predictor variables, like EL or SED status, may improve the accuracy of screening (Johnson et al., 2010).

There are some other limitations to CBM-R (Wayman et al., 2007). For instance, it may not be an appropriate measure for beginning readers. When the first graders were tested for oral reading fluency, a floor effect resulted (Fuchs, Fuchs & Compton, 2004). Word identification measures proved to have more promising results for readers at the beginning stages (Wayman et al., 2007). Compton and colleagues (2006) studied the use of a word identification measure to identify at-risk students in the first grade. The performance level and slope were measured to predict performance at the end of second grade. Compton et al. found that performance level and slope of a word identification probe given to first graders considerably improved the classification accuracy of at-risk students over all other CBM measures of reading.

Wayman et al. (2007) offered some words of caution that mobilized researchers to investigate different aspects of CBM. First, despite the consistent correlations of .70 and higher between CBM-R and other criterion measures, there are some studies with correlation as low as .40. Some of these studies indicated that the relationship between CBM-R and criterion variables might decrease for older students. That is, correlations tended to drop at the intermediate grades. This decline in the upper grades suggests that CBM-R may not be the best measure to be used with older students. Instead, a multivariate approach that uses multiple measures and other
predictor variables, like EL or SED status, may improve the accuracy of screening (Johnson et al., 2010). Other evidence suggested that this relationship could be influenced by different factors, such as passage characteristics. Second, because most of the research has been done on CBM-R in grades 2-5, it calls into question the validity of CBM-R for K-1 students and older students in grades 6-12. Finally, with the broader application of CBM, especially when used to make higher-stakes decisions, more empirical evidence is needed regarding the technical adequacy of slopes produced by CBM measures (Wayman et al., 2007).

Since Wayman's et al., (2007) review, researchers have continued to investigate the correlational relationship between CBM-R and state standardized tests and other norm-referenced standardized measures of general reading ability and found moderate to strong correlations (Ditkowski & Koone, 2009; Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008; Valencia et al., 2010; Yeo, 2011). On the other hand, some researchers have recently expressed concerns about using a GOM to track and evaluate individual student's progress and response to intervention (Ball & Christ, 2012). The level at which schools are struggling the most with making data-based decisions is at the individual student level (Ball & Christ). A major concern cited is the overgeneralization and interpretation of CBM-R as valid and reliable for evaluation and progress monitoring purposes due to its existing evidence and reputation for benchmarking and universal screening (Ardoin, Christ, Morena, Cormier, & Klingbeil, 2013). Ardoin and colleagues (2103) completed an extensive review of the literature related to progress monitoring and decision rules. They concluded that there is a lack of psychometric and empirical evidence to support these approaches (i.e., data point rule, trend line rule).

Despite over thirty years of research supporting the technical adequacy of CBM, there remain large gaps in research and practice regarding proper administration, instrumentation, and
uses of CBM within a RtI Framework in general, and especially at the middle and high school levels (Tichá, Espin, & Wayman, 2009). More research is needed to determine the technical adequacy of CBM to be used for placement and intervention decisions beyond reliability and validity (Ardoin & Christ, 2009; Christ & Ardoin, 2009; Poncy, Skinner, & Axtell, 2005). Furthermore, research is needed to investigate the accuracy of other screening measures used in identifying student difficulties in content areas other than reading (Glover & Albers, 2007).

**Computer Adaptive Tests (CATs)**

CATs are an alternative approach of universal screening that is typically group-administered, and computer-scored. The notable feature of the CAT design is the adaption of the test taker's responses; when a student answers a question correctly, a more difficult item is displayed, and when a student answers incorrectly, an easier item is given (Moyer, Galindo, & Dodd, 2012). STAR (https://www.renaissance.com) and MAP (https://www.renaissance.com) are examples of CATs. A distinguished hallmark of CATs is that scores are reported on an equal interval vertically equated Rasch Unit (RIT) scale, which allows for comparisons of student growth within and across grade levels.

At this time, there is minimal evidence cited in the literature that has examined the utility of any CAT in predicting performance on an outcome reading measure. Several unpublished dissertations (e.g., Andren, 2010; Curry, 2016; Jones, 2015; Maziarz, 2010) reported statistically significant correlations between MAP and state test scores. Other studies have investigated the relationship between MAP-R and GOMs (e.g., January & Ardoin, 2015; Merino & Beckman, 2010) or the predictive utility of combined MAP-R and GOMs scores (Ball & O'Connor, 2016; Klingbeil, McComas, Burns, & Helman, 2015). Van Norman and colleagues (2017) examined the use of a CAT (i.e., Classworks Reading) and the previous year's test scores to predict the
current year state test outcome and determined the CAT’s given cut score for Grades 5 through 8 had acceptable sensitivity (ranges from .77 to .92) to use as a universal screener. January and Ardoin (2015) conducted a study on third-graders and reported initial evidence of the concurrent validity of MAP-R and a standardized state test. In another study, Klingbeil et al. (2017) examined three approaches (i.e., CAT, CBM, running records) to universal screening and found that MAP-R demonstrated the most promise as a single predictor of performance on a summative reading assessment in Grades 3 through 5.

**Summary of Universal Screening**

Despite the promising research showing the technical adequacy of CATs like MAP-R, most research evidence suggests that combining multiple screening measures improves the diagnostic accuracy (Compton et al., 2006; Johnson et al., 2010; O'Connor & Jenkins, 1999; Shapiro et al., 2008). Speece and et al. (2010) found that batteries with two to three measures outperformed, batteries with more measures. Furthermore, the use of multiple tests can create economic and time restraints that are costly for districts. The evidence is sparse; however, regarding a single measure as a technically adequate screening measure (Johnson et al., 2009; Ritchey & Speece, 2004).

Additionally, the majority of the research on universal screeners for reading has focused on students in the elementary grades using GOMs like CBM. Klingbeil and colleagues (2017) found nine screening measure that examined the diagnostic accuracy of universal screening measures in the upper elementary or middle-level grades (i.e., Baker et al., 2015; Decker et al., 2014; Denton et al., 2011; McGlinchey & Hixson, 2004; Nese, Park, Alonzo, & Tindal, 2011; Shapiro et al., 2008; Speece et al., 2010; Stage & Jacobsen, 2001; Stevenson, Reed, & Tighe, 2016). Several notable trends emerged from this body of research. For example, none of the
outcome measures used were aligned to the Common Core State Reading Standards. Another key finding was that researchers used the publisher-provided cut scores instead of local norms or statistically derived cut scores and when using these cut scores with single screening measures, diagnostic accuracy results tended to be below the level of acceptability for a universal screener.

Quick and accurate identification of struggling readers in middle grades is essential for the effectiveness of multi-tiered service delivery models like RtI (Glover & Albers, 2007). As such, decisions made policymakers, curriculum leaders, and practitioners should not be taken lightly. Educators must decide which reading skills to assess, how to assess those skills, and with what and how many measures to use. From there, school teams must decide how to use the results to identify struggling readers and determine what kind of intervention is needed. There is limited empirical research, however, to guide middle-school teams in making procedural decisions around universal screening. A widely used CAT, Measures of Academic Progress for Reading (MAP-R), shows promise as a single screening measure used for universal screening, thus, is the focus of this study.
Chapter 3

METHOD

Universal screening assessments help education practitioners identify students who are at risk in reading and need interventions and supports. For this study, I examined the technical adequacy of the Measures of Academic Progress for Reading (MAP-R; Northwest Evaluation Association [NWEA], 2011; 2013; 2015a) and demographic variables in predicting student performance on the eMPowerME English Language Arts/Literacy (eMPowerME ELA/L; Measured Progress, 2016; 2017). I also determined the minimum MAP-R scores needed in the spring of Grade 6 and 7 that demonstrated a student was on track to meet proficiency standards on the eMPowerME one year later in Grade 7 and 8. A secondary purpose was to determine whether differences existed between middle-level students’ subtests scores on the MAP-R. The investigation addressed the following research questions:

1. What is the concurrent validity of the MAP-R and the eMPowerME ELA/L in Grades 7 and 8?

2. How useful are MAP-R scores in identifying sixth and seventh graders who will and will not meet proficiency standards on eMPowerME ELA/L one year later?

3. Is the predictive validity of the MAP-R scores improved with the inclusion of demographic variables (i.e., gender, SES, disability status)?

4. What are the optimal cut scores on MAP-R in spring of Grades 6 and 7 for predicting performance on the eMPowerME ELA/L in Grades 7 and 8?
5. Are there differences between and among subtest scores on the MAP-R for sixth, seventh, and eighth graders, and if so, do differences vary according to student demographics (i.e., gender, SES, disability status)?

6. What unique contribution of MAP-R literary text and informational text in explaining variance in performance in eMPowerME ELA/L in sixth-seventh and seventh-eighth graders?

In this chapter, I discuss the design and nature of the study, including the methods employed to recruit schools and obtain data files, the study sample and measures used, and data analysis procedures. I used SPSS version 22 for all analyses.

**Procedure for Data Collection**

In May 2017, I took several steps to select schools and obtain the necessary data to complete this study. Using the Google search engine and the terms *Maine, NWEA MAP, RSU,* and *middle school,* I identified 25 districts around the state of Maine that indicated on their websites that they used NWEA’s tests. My initial goal was to include three schools with databases that included spring 2015 MAP-R and Spring 2016 eMPowerME ELA/L student data for Grades 6 through 8 and student demographic data (i.e., grade level, gender, SES, disability status). I narrowed the list to six schools and located the contact information of district personnel responsible for warehousing student testing and demographic information. I selected three of the six schools because I had a personal connection with an employee in these districts. All six school locations were close to me.

My method of the first contact with the schools was via phone call. I introduced myself, provided a brief overview of my study, explained why I chose that particular school, and described data files needed to do my research. I confirmed that the school included Grades 6
through 8 and had MAP-R scores for spring 2015 and 2016 testing periods. If the school representative was unable to answer my questions, he or she provided the name and contact information of the person with knowledge of available data. In each case, the school employee expressed an interest in working with me. I followed up with a formal email thanking that person for agreeing to move forward with determining what would be required to create the data files and get approval from his or her supervisor (i.e., building principal, assistant superintendent, superintendent) to release this information to me. Additionally, I sent a formal email to the school/district administrator introducing myself and included a recap of my initial conversation with the representative, a brief overview of my study and IRB status, and an explanation of the required data needed to conduct my analyses. I attached to the email a pdf file of my study proposal if he or she wanted to review further details about my proposed research.

Two schools immediately declined to participate in the study. After several communications back and forth, a third school discontinued their involvement in mid-August 2017. A fourth school withdrew from their commitment at the end of December 2017 after several months of attempting to gather and organize the data. All four schools’ spokespersons cited lack of resources as the reason for not participating, including recent changes in data management personnel, practices, and systems. These schools did not have a universal data management system for storing student test scores, nor did they have the human resources to compile the data at that time. The sample for the present study, therefore, was limited to two middle schools, identified as A and B, with established data systems. I secured permission to obtain demographic information and testing data from the two schools in June 2017. I received the first round of data from school A at the end of June 2017 and school B by the end of July 2017.
Each district representative reported that personnel in their school had administered assessment measures according to standardized procedures. Scores for the MAP-R and eMPowerME ELA/L were computer generated and entered into data files by each district. To ensure the anonymity of individual test data, the data coordinators de-identified data files before releasing information to me. Given that these were archival data files with non-identifiable student data, I was not required to obtain parental consent to include a student’s score in my study. The University of Maine’s Institutional Review Board reviewed and approved this project in May of 2017 under the methods and procedures applied to the use of archival data.

Upon reviewing the data files, I identified a critical missing data issue. School A did not have spring MAP-R test scores for the Grade 5-6 cohort but did for the Grades 6-7 and 7-8 cohorts. Second, school B inadvertently included spring 2016 MAP-R scores for all three cohorts instead of spring 2015. I contacted the representative from school A and learned that school A no longer tested fifth graders using MAP-R. For this reason, I eliminated the Grade 5-6 cohort from the sample, leaving two grade-based cohorts, a Grades 6-7 cohort who were Grade 6 in 2015-16 and Grade 7 in 2016-17, and a Grades 7-8 cohort who were in Grade 7 in 2015-16 and Grade 8 in 2016-17. Also, given the time that had elapsed between my initial recruitment effort and receipt of the data, spring 2017 eMPowerME ELA/L scores were now available. I, therefore, opted to use more recent eMPowerME scores from spring 2017 rather than spring 2016 and to investigate the predictive validity of the MAP-R using spring 2016 instead of spring 2015 scores and concurrent validity using spring 2017 scores.

**Sample Description**

For this study, I used existing data from student cohorts in Grades 6-7 and 7-8 from two middle schools in public districts located in rural Western Maine who had completed (a) the
MAP-R screening in both or either spring 2016 and spring 2017 and (b) the eMPowerME ELA/L in spring 2017. Both schools served Grades 6 through 8. The data files contained demographic information for all students participating in the study, including grade level, gender, SES, and disability status. Given that students who are male (Klecker, 2006, 2014; Reardon, Fahle, Kalogrides, Podolsky, & Zárate, 2018), students from low-income households (Hernandez, 2011; Lacour & Tissington, 2011; Rampey et al., 2009; Reardon, 2013), and students with learning disabilities are at higher risk of dropout and underperformance in reading, it is vital that screening instruments work as equally well at identifying specific subgroups of students as with the general population.

Table 1 depicts the demographics at the time of the October 2016 enrollment count for each school, according to the Maine DOE website. Descriptively, the schools enrolled slightly more males (50.96%) than females, 48.91% of students received free or reduced-price lunch (FRL), and 19.38% of students had Individualized Education Plans (IEPs). The Maine DOE enrollment data indicated that 89.36% of all public-school students PK-12 were White at the April 2017 count. The percentage of White students attending the schools included in this study was 96.66% and 97.17% for schools A and B respectively. Given the overall lack of racial diversity in the state and the participating schools, I did not examine race as a variable.

According to the Maine Assessment and Accountability Reporting System (MAARS) portal on the Maine DOE website, 97.51% of eligible Maine students in Grades 3 through 8 participated in the spring 2017 administration of the eMPowerME ELA/L test. The two schools in this study had participation rates that were comparable to the overall state average for all grades tested (96.39% for School A and 98.21% for School B). The site also reported that 52.56% of all Maine students scored at or above proficiency based on state standards.
Descriptively, the percentage of students at or above proficiency in my two schools was lower than the statewide rate, particularly for school B (School A = 41.99% and School B = 28.84%).

Table 1

*School Demographic Data (Grades 6, 7, 8)*

<table>
<thead>
<tr>
<th>School</th>
<th>Male n (%)</th>
<th>Female n (%)</th>
<th>No FRL n (%)</th>
<th>FRL n (%)</th>
<th>No IEP n (%)</th>
<th>IEP n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>255 (50.80)</td>
<td>247 (49.20)</td>
<td>271 (53.98)</td>
<td>231 (46.02)</td>
<td>392 (78.09)</td>
<td>110 (21.91)</td>
</tr>
<tr>
<td>B</td>
<td>142 (51.26)</td>
<td>135 (48.74)</td>
<td>127 (45.85)</td>
<td>150 (54.15)</td>
<td>236 (85.20)</td>
<td>41 (14.80)</td>
</tr>
<tr>
<td>Total</td>
<td>397 (50.96)</td>
<td>382 (49.04)</td>
<td>398 (51.09)</td>
<td>381 (48.91)</td>
<td>628 (80.62)</td>
<td>151 (19.38)</td>
</tr>
</tbody>
</table>

*Note:* Maine DOE Education Data, October 2016 enrollment count. FRL = free/reduced price lunch; IEP = Individualized Education Program

**Measures**

**Measures of Academic Progress-Reading (MAP-R)**

The computer-adaptive MAP-R assessment by Northwest Evaluation Association (NWEA) for sixth, seventh, and eighth grades served as the independent variable in my study due to the use of the test at these grade levels and ability to compare scores on reading subtests within and across grade levels. Almost eight million students complete the MAP tests annually as a universal screening/benchmarking measure (NWEA, 2011; 2013; 2015a; 2015b). A computer-adaptive assessment allows for a more accurate evaluation of mastery using fewer questions as compared to conventional testing procedures (Weiss & Kingsbury, 1984). The computer-adaptive MAP tests begin with a question appropriate for the student’s grade level, then adjusts throughout the test in response to student performance; a correct answer generates a more difficult question, whereas an incorrect answer generates a less difficult question. MAP tests were not timed; however, the approximated time to take each assessment was between 50
and 60 minutes (NWEA, 2013; 2015a). Generic NWEA documentation assumes the default weeks of instruction: 4, 20, and 32 weeks of instruction for fall, winter and spring testing respectively, and 36 weeks for the next academic year.

The typical reading assessment contained 42 questions that NWEA scored using Rasch Units abbreviated as RIT (NWEA, 2015a). NWEA defined RIT scales as equal-interval scales that allowed for measurement over time regardless of grade level or age of the student. The MAP-R provides an overall composite score and scores for three goal areas: (a) literature, (b) informational text, and (c) vocabulary acquisition and use. In 2015, the NWEA published a technical document that identified MAP cut-scores that corresponded to the Smarter Balanced summative assessment achievement levels. In addition to RIT scores, quintile bands were reported based on MAP-R cut-scores with the following descriptors: (1) low meant student performance was below the 20th percentile, (2) low-average indicated performance between the 21st and 40th percentile, (3) average meant performance between the 41st and 60th percentile, (4) high-average depicted performance between the 61st and 80th percentile, and (5) high represented student performance above the 81st percentile. Periodically, NWEA has conducted norming studies with the most recent study by Thum and Hauser (2015) that evaluated nearly 500,000 tests scores for over 100,000 students. The 2015 RIT scale norms were developed using nine datasets from fall 2011 through spring 2014 (NWEA, 2015a).

According to the technical manual, the adaptive nature of the MAP-R requires alternative approaches for determining test-retest reliability, parallel forms reliability, and internal consistency (NWEA, 2011). Specifically, the manual reported reliability based on a combination of test-retest and a type of parallel forms reliability called stratified, randomly-parallel form reliability (Green, Bock, Humphreys, Linn, & Reckase as cited in NWEA, 2011). The timeframe
between test administrations was a few months, and the second test was not the same form. Reliability was examined based on correlations between (a) "two tests administered from two different but related item pools and those administered twice but from different item pools" and (b) "scores from tests taken in one term (e.g., spring or fall) with the same students tested the following fall or spring term" (NWEA, 2011, p. 55). All reliability coefficients reported in the technical manual for reading composite scores in Grades 6 through 8 ranged from .74 to .84. Internal consistency reliabilities of MAP-R composite scores consistently were in the low to mid .90s for Grades 6 through 8. Reliabilities for reading subscores were not reported.

NWEA conducts regular linking studies to compare MAP-R scores and state standardized tests to measure student achievement. According to the NWEA website (nwea.org), recent studies (February 2016 - June 2018) examining the concurrent validity of MAP-R composite scores in Grades 6 through 8 have been investigated using state assessment results in 28 states, not including Maine. Correlations between MAP-R and state assessment results have ranged from .72 to .89. In 2010, NWEA conducted a study of the alignment of the MAP-R RIT scale with the New England Common Assessment (NECAP). Along with New Hampshire, Rhode Island, and Vermont, Maine used the NECAP as the state assessment from 2009 to 2014; however, the data used in the alignment study were derived from the New Hampshire students only.

**eMPowerME English Language Arts/Literacy (ELA/L)**

I used the computerized comprehensive measure of reading skills, eMPowerME ELA/L by Measured Progress as the dependent variable in this study because it is the test connected to accountability for school districts in the state of Maine. According to the technical report, Measured Progress designed eMPowerME to measure academic content standards in
mathematics and ELA/Literacy (Measured Progress, 2017). The state updated the Maine Learning Results (MLR) in 2011 to include the Common Core State Standards (CCSS) as the college and career readiness standards. School districts first administered the assessment to 78,000 public school students in Grades 3 through 8 via standard administration or administration with accommodations during March–April 2016 after the official adoption of the test in December 2015. Reports included student results according to academic achievement descriptors utilizing cut scaled scores established using standard-setting procedures for each of four levels of achievement with concerning to state expectations. Table 2 includes the achievement level descriptors and range of scores for Grades 7 and 8.

Student reports expressed scores on the 2016-17 eMPowerME tests in raw or scaled forms. Similar to the NWEA MAP, because the assessment has the same vertical scale for each subject area, educators can assess growth as students move from one year to the next. The ELA/Literacy score is comprised of the reading, language, and writing subtests (Measured Progress, 2017). The reporting categories for reading include comprehension, analysis, and interpretation for both focus areas of literature and informational text. The language test covers revising informational text and argument text and English language and conventions (including vocabulary acquisition and use). Although schools in the present study might have received scores in their district reports in all these reporting categories, they did not extract them to include in their databases. For this reason, I obtained only eMPowerME (composite) ELA/L scores from the schools.

According to the technical report (Measured Progress, 2017), the split-half method using Cronbach’s alpha was used to assess the reliability of the 2016–17 eMPowerME tests. Cronbach’s alpha coefficients ranged from 0.90 to 0.91 for Grades 6, 7, and 8 with Standard
Error of Measurement (SEM) of raw scores ranging from 3.56 to 3.66. Double-blind scoring was employed to monitor the quality of the hand-scoring of student responses for constructed-response items. Interrater consistency correlations ranged from 0.67 to 0.78. Overall decision accuracy ranged from 0.68 to 0.71 with consistency ranges from 0.77 to 0.79 (Kappa ranges = 0.56 to 0.59).

Table 2

*eMPowerME English Language Arts/Literacy Achievement Level Descriptors and Score Ranges*

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Score Ranges</th>
<th>Grade 7</th>
<th>Grade 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 Above State Expectations</td>
<td>The student’s work demonstrates a <strong>thorough</strong> understanding of the knowledge and skills needed to meet Maine’s ELA/Literacy Content Standards with text at the appropriate complexity for the grade level.</td>
<td>776-790</td>
<td>878-890</td>
</tr>
<tr>
<td>3 At State Expectations</td>
<td>The student’s work demonstrates an <strong>adequate</strong> understanding of the knowledge and skills needed to meet Maine’s ELA/Literacy Content Standards with text at the appropriate complexity for the grade level.</td>
<td>760-775</td>
<td>860-877</td>
</tr>
<tr>
<td>2 Below State Expectations</td>
<td>The student’s work demonstrates an <strong>incomplete</strong> understanding of the knowledge and skills needed to meet Maine’s ELA/Literacy Content Standards with text at the appropriate complexity for the grade level.</td>
<td>745-759</td>
<td>844-859</td>
</tr>
<tr>
<td>1 Well Below State Expectations</td>
<td>The student’s work demonstrates a <strong>minimal</strong> understanding of the knowledge and skills needed to meet Maine’s ELA/Literacy Content Standards with text at the appropriate complexity for the grade level.</td>
<td>700-744</td>
<td>800-843</td>
</tr>
</tbody>
</table>

*Source:* Maine DOE Website October 2018
Procedure for Data Analysis

To inform the first two research questions, I computed a Pearson product-moment correlation coefficient to measure the strength and direction of the linear relationship between scores on the spring 2016 and 2017 MAP-R and scores on spring 2017 eMPowerME ELA/L tests for both Grade 6-7 and 7-8 cohorts. The value can range from -1 for a perfect negative linear relationship to a +1 for a perfect positive linear relationship (Cohen, 1988). These correlations provided opportunities to understand the magnitude of the relationship between the variables better, and also served as the basis for additional analyses.

With the primary outcome of consideration being accuracy in predicting proficiency on the eMPowerME ELA/L, the result of greatest interest is dichotomous. For this investigation, a student meets or does not meet proficiency on the eMPowerME ELA/L test served as the dependent variable. The predictive factors that were analyzed included the continuous variables of the MAP-R Rasch Unit (RIT) score, and the demographic variables of gender, SES, and disability status served as dichotomous predictor variables. Gender indicates the physical attributes used to assign a student's sex (male/female). SES references the economic status of the student's family based on free or reduced-price lunch status (No FRL/Yes, FRL). Disability denotes whether or not a student has an Individualized Education Plan (IEP; No IEP/Yes IEP).

For testing of categorical and continuous predictors on dichotomous outcomes, binomial logistic regression is the most appropriate method of analysis (Peng, Lee, & Ingersoll, 2002; Silberglitt & Hintze, 2005). To address the second and third questions, I conducted two regression analyses to determine if MAP-R scores as a solo predictor or MAP-R scores combined with demographic variables added value when assessing sixth and seventh graders performance (proficient–not proficient) one year later on eMPowerME ELA/L. These analyses
provided specificity (SEN) and sensitivity (SPE) levels, as well as the overall percentage accuracy in classification or classification accuracy (CA).

Although, the classification accuracy of an individual or combined predictor(s) is essential to determine, ultimately, the primary goal of any educational screening process is to maximize the accurate identification of those students who require intervention. Thus, further data analyses were needed to identify specific cut points associated with desired sensitivity levels. I used receiver operating characteristic (ROC) curve analyses to respond to the fourth question regarding the optimal cut scores on spring MAP-R in Grades 6 and 7 to predict performance on the eMPowerME ELA/L in Grades 7 and 8. ROC involves plotting the true-positive rate (sensitivity) on the y-axis and the false-positive rate (1-specificity) on the x-axis for each possible screening cut score (Hilbe, 2009). The area under the curve (AUC) describes the overall accuracy of the screener with values ranging from .5 to 1.0, where higher values indicate higher classification accuracy. According to Hosmer et al. (2013), an AUC between .50 and .70 means poor accuracy, between .70 and .80 is good, between .80 and .90 is excellent, and higher than .90 is considered an outstanding level of accuracy.

I used a one-way repeated measures analysis of variance (ANOVA) and independent-samples t-tests to answer the fifth question on whether there were any significant differences between student scores on the three MAP-R subtests and if there were any significant differences among subtests according to student characteristics. A one-way repeated measures ANOVA is the best analysis for this research question because the same participants are being measured on the same dependent variable, reading comprehension, but under three different conditions (i.e., literary text, informational text, vocabulary acquisition and use). The SPSS output indicated whether an overall significant difference existed in the mean scores of the three subtests, but did
not point to exactly where the difference occurred between subtests. Thus, the output was augmented to include a Bonferroni post hoc test, which is recommended by Maxwell and Delaney (2004) for testing all pairwise combinations of levels of the within-subjects factor. The independent-samples $t$-test is the analysis to determine if a statistically significant difference exists between the means of two independent groups on a continuous dependent variable.

Finally, to examine the contributions of MAP-R literary text and informational text subtest scores in explaining the unique variance in performance on the eMPowerME ELA/L, I conducted a hierarchical logistic regression. I entered vocabulary acquisition and use in block 1, literary text in block 2, and informational text in block 3 of the analyses.

**Summary**

The purpose of this chapter was to report an overview of how the study was conducted. I described the research questions and an overview of the study, participants, and procedures for data analysis. Pearson product-moment correlation coefficients, (hierarchical) binomial logistic regression, and one-way repeated measures ANOVA analyses were well suited for a study of this nature to determine the effectiveness of the MAP assessment, in concert with various demographic factors, in predicting eMPowerME ELA/L assessment passing rates on seventh- and eighth-grade reading scores. The study was intended to examine relationships among variables. I provide the results of the data analysis in Chapter 4.
Chapter 4

RESULTS

The primary purpose of this study was to add to the literature on the relationship between a universal screening measure and a statewide summative reading achievement test. Specifically, I examined the concurrent and predictive validity of the Measures of Academic Progress for Reading (MAP-R; NWEA, 2011, 2013, 2015a) with Maine’s eMPowerME English Language Arts/Literacy (ELA/L) test for students in the sixth, seventh, and eighth grade. Additionally, I used gender, socioeconomic status (SES), and disability status to determine if there was value added by using the MAP-R scores along with certain student characteristics to predict performance on the eMPowerME ELA/L. In this investigation, I also determined the minimum MAP-R scores needed in the spring of Grade 6 and 7 that demonstrated a student was on track to meet proficiency standards on the eMPowerME one year later in Grade 7 and 8. My secondary purpose of the study was to examine whether or not there were considerable differences between middle-level students' performance on the MAP-R subtests (i.e., literary text, informational text, vocabulary acquisition and usage) as well as if there were any differences within subtest scores by demographic variables. Finally, I examined the unique contributions of literary text and informational text comprehension (after controlling for vocabulary acquisition and use) in explaining variance in eMPowerME ELA/L performance for students in sixth-seventh and seventh-eighth grades. In this study, I addressed the following research questions:

1. What is the concurrent validity of the MAP-R and the eMPowerME ELA/L in Grades 7 and 8?

2. How useful are MAP-R scores in identifying sixth and seventh graders who
will and will not meet proficiency standards on eMPowerME ELA/L one year later?

3. Is the predictive validity of the MAP-R scores improved with the inclusion of demographic variables (i.e., gender, SES, disability status)?

4. What are the optimal cut scores for MAP-R in spring of Grades 6 and 7 for predicting performance on the eMPowerME ELA/L in Grades 7 and 8?

5. Are there differences between and among subtest scores on the MAP-R for sixth, seventh, and eighth graders, and if so, do differences vary according to student demographics (i.e., gender, SES, disability status)?

6. Do MAP-R literary text and informational text scores provide unique contributions in explaining variance in eMPowerME ELA/L performance for sixth-seventh and seventh-eighth graders?

In the subsequent sections, I outlined the characteristics of my sample, including the MAP-R participation rates by grade-cohort, demographics, and eMPowerME ELA/L achievement levels. Next, I described the results of analyses that examined the relationship between the performance on MAP-R and eMPowerME ELA/L taken concurrently and across one year and whether or not specific student characteristics added to the strength of the association. Additionally, I examined the MAP-R cut scores that maximized sensitivity and specificity. In the final section of the chapter, I examined analyses that compared student performance between and among the three MAP-R subtests as well as the contributions of each MAP-R subtest in explaining unique variance in performance on the eMPowerME ELA/L.
Descriptive Statistics

Description of the Sample

The initial sample for this study included 484 students in Grades 6-7 and 7-8 from two schools, A (n = 319) and B (n = 165) who had participated in the ELA/L component of the spring 2017 eMPowerME (Measured Progress, 2016; 2017). According to district data personnel, the small percentage (2%) of students in each school who did not participate in the assessment were either (a) individuals with disabilities who participated in the Maine Educational Alternative Assessment, the Multi-State Alternate Assessment (MSAA), (b) students whose parents opted out of state and local testing, or (c) those who were not attending the school at the time of the test. The sample comprised of an equal proportion of students at each grade level (Grades 7 = 49.79% and Grade 8 = 50.21%). Descriptively, there were more male (52.47%) than female (47.52%) participants, and a greater percentage of students received FRL (48.55%) or had an IEP (21.28%) than was the case statewide at those grade levels (42.63% and 18.29%, respectively) (Maine DOE Education Data, April 2017).

I also compared the achievement level on eMPowerME ELA/L of all students in the sample to statewide data to determine how well my sample represented seventh and eighth graders in Maine. On eMPowerME, four levels of performance were possible: level 1 (well-below state expectations), 2 (below state expectations), 3 (at state expectations, and 4 (above state expectations). According to the 2016-2017 eMPowerME ELA/Literacy & Mathematics Technical Report, in spring 2017, 52.23% of Grade 7 and 51.52% of Grade 8 students identified as proficient on the ELA/L portion of the test. These students scored at either a level 3 or level 4. In the present study, 38.30% of Grade 7 and 39.50% of Grade 8 students in the sample met proficiency standards (i.e., scores at level 3 or 4) on the spring 2017 eMPowerME ELA/L, a
proficiency rate that was lower than the rate statewide for those grades (52.23% and 51.52%, respectively).

**Participation in MAP-R by Grade Cohort.** A primary purpose of the present study was to investigate the validity of MAP-R (Thum & Hauser, 2015) scores in identifying students who were and were not likely to be successful in meeting proficiency standards on the eMPowerME ELA/L. Not all students in the sample, however, participated in the MAP-R test in the two years for which I requested scores, spring 2016 and spring 2017. For this reason, I investigated whether there were significant differences in the rate of MAP-R participation for students by grade cohort, student characteristics, and achievement on the eMPowerME ELA/L test using the chi-square test of independence. This test determines whether two variables are statistically independent. I identified four MAP-R participation groups: students who had scores for (a) both spring 2016 and 2017 (n = 414), (b) spring 2016 only (n = 25), (c) spring 2017 only (n =34), and (d) neither spring 2016 or 2017 (n = 11).

Table 3 displays the distribution of MAP-R participation by grade cohort. As can be seen, the vast majority of students (85.54%) had participated in the MAP-R in both the spring 2016 and spring 2017, and among the 14.41% who did not have scores for both years, only 2.27% had scores for neither year. Although the MAP-R participation rate was fairly similar in the Grade 6-7 and 7-8 cohorts, the results of chi-square analyses indicated that students in the 7-8 cohort were somewhat over-represented in the group who took the MAP-R in spring 2016 but not in spring 2017, $\chi^2(3, n = 484) = 9.44, p = .024$.

**Participation in MAP-R by Student Demographics.** As seen in Table 4, I also examined the key characteristics of students by participation group to determine whether there were systematic differences in group composition. Descriptively, there was some variability in gender, SES, and
disability status across MAP-R participation groups. The results of chi-square analyses showed that the pattern of MAP-R participation did not vary by gender, $\chi^2 (3, n = 484) = 2.44, p = .49$, but students who received free and reduced-price lunch were somewhat over-represented in the group who did not take the MAP-R in spring 2017 or either 2016 or 2017, $\chi^2 (3, n = 484) = 17.5, p = .001$. In addition, students with an IEP were over-represented in the group who did not take the MAP-R in either spring 2016 or 2017, $\chi^2 (3, n = 484) = 25.12, p < .001$.

**Participation in MAP-R by eMPowerME ELA/L Achievement Levels.** Next, I considered whether there were differences in achievement levels among students who varied in years of MAP-R participation. Specifically, I compared MAP-R participation rates for students who did and did not meet proficiency standards on the eMPowerME ELA/L in spring 2017. As I mentioned earlier, on eMPowerME, four levels of performance were possible: level 1 and 2 denote well-below and below state expectations and levels 3 and 4 represent at and above state expectations. Using chi-square analysis, I compared years of participation in MAP-R for students who met (i.e., level 3 or 4) and did not meet (i.e., level 1 or 2) the standard for proficiency on eMPowerME ELA/L in spring 2017. The results indicated no significant difference in proficiency by MAP-R participation pattern, $\chi^2 (3, n = 484) = 5.32, p = .150$. As seen in Table 5, the group who did not take the MAP-R in either year was least likely to be proficient, and the group who only took the MAP-R in spring 2016 had a slightly higher rate of proficiency than all three other groups. The groups who took MAP-R in both years and spring 2017 only were similar to the percentage of students who were proficient on eMPowerME ELA/L.
Table 3

**MAP-R Participation Rates by Grade Cohort**

<table>
<thead>
<tr>
<th>Grade Cohort</th>
<th>MAP-R Participation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n (%)</td>
</tr>
<tr>
<td>Spring 16/17</td>
<td>212 (88.0)</td>
</tr>
<tr>
<td>Spring 16</td>
<td>5 (2.1)</td>
</tr>
<tr>
<td>Spring 17</td>
<td>18 (7.4)</td>
</tr>
<tr>
<td>No Spring 16 or 17</td>
<td>6 (2.5)</td>
</tr>
<tr>
<td>Total</td>
<td>414 (85.5)</td>
</tr>
</tbody>
</table>

*Note: MAP-R = Measures of Academic Progress-Reading.*

Table 4

**MAP-R Participation Rates by Student Demographics**

<table>
<thead>
<tr>
<th>Student Characteristics</th>
<th>MAP-R Participation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n (%)</td>
</tr>
<tr>
<td>Spring 16/17</td>
<td>219 (52.9)</td>
</tr>
<tr>
<td>Spring 16</td>
<td>14 (56.0)</td>
</tr>
<tr>
<td>Spring 17</td>
<td>14 (41.2)</td>
</tr>
<tr>
<td>No Spring 16 or 17</td>
<td>7 (63.6)</td>
</tr>
<tr>
<td>Total</td>
<td>254 (52.5)</td>
</tr>
</tbody>
</table>

*Note: FRL = free/reduced price lunch; IEP = Individualized Education Program.*

Table 5

**MAP-R Participation Rates by eMPowerME ELA/L Achievement Levels**

<table>
<thead>
<tr>
<th>Achievement Level</th>
<th>MAP-R Participation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n (%)</td>
</tr>
<tr>
<td>Spring 16/17</td>
<td>256 (61.8)</td>
</tr>
<tr>
<td>Spring 16</td>
<td>11 (44.0)</td>
</tr>
<tr>
<td>Spring 17</td>
<td>22 (64.7)</td>
</tr>
<tr>
<td>No Spring 16 or 17</td>
<td>9 (81.8)</td>
</tr>
<tr>
<td>Total</td>
<td>298 (61.6)</td>
</tr>
</tbody>
</table>

*Note: MAP-R = Measures of Academic Progress-Reading.*
Given the necessity of having MAP-R scores to address my research questions, students who were missing both 2016 and 2017 MAP-R scores (n = 11) were excluded from all subsequent analyses, resulting in a final sample of 473 students, 235 in the Grade 6-7 cohort and 238 in the Grade 7-8 cohort. A total of 439 students had both Spring 2017 eMPowerME ELA/L and spring 2016 MAP-R data (i.e., Grade 6-7 cohort = 217; Grade 7-8 cohort = 222) and 448 had both Spring 2017 eMPowerME ELA/L and spring 2017 MAP-R data (i.e., Grade 6-7 cohort = 230; Grade 7-8 cohort = 218). Although only a small number of students did not take the MAP-R in either year, as I mentioned above, these excluded students were more likely to have received free or reduced-price lunch or to have an IEP. This issue will be discussed further in chapter 5. Students who were missing just one year of MAP-R participation, spring 2016 or spring 2017, were retained in analyses involving the year in which they participated.

**Performance on MAP-R and eMPowerME ELA/L.** Table 6 shows the breakdown of the number and percentage of students who were and were not proficient on the state test according to MAP-R participation in the sample and the overall state participation sample. The MAP-R tests reported student performance on a single scale using Rasch Units abbreviated as RIT (NWEA, 2015a). Data files that schools obtained from MAP-R included both a total RIT score for reading and RIT scores disaggregated by three content goals (literature, informational text, and vocabulary acquisition and use). As I explained in chapter 3, in addition to RIT scores, quintile bands were reported based on MAP-R national norms with the five descriptor levels. To determine the criteria for proficiency for the total reading RIT score on the MAP-R, I used the established MAP-R norms for Grades 6, 7, and 8 (Thum & Hauser, 2015) and identified student scores that were at or below the 40th percentile (i.e., low, low-average) as not proficient and above the 40th percentile (i.e., average, high-average, high) as proficient. School data files also
included two types of eMPowerME EL/L scores: scaled scores and achievement-level descriptors. Scaled scores ranged from 700 to 790 for Grade 7 and 800 to 890 for Grade 8. Items on the eMPowerME ELA/L tests were developed to assess the Maine Learning Results (MLR) based on the Common Core State Standards for English Language Arts (CCSS-ELA). The schools’ data reports organized student scores at the *well-below* and *below state expectations*, as not proficient, and scores *at* and *above state expectations* as proficient.

Table 6

*MAP-R Participation by Grade Cohort According to eMPowerME-ELA/L Achievement Levels*

<table>
<thead>
<tr>
<th>Grade Cohort</th>
<th>6-7</th>
<th>7-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>eMPowerME-ELA/L Achievement Level</td>
<td>Not Proficient</td>
<td>Proficient</td>
</tr>
<tr>
<td>Spring 2016</td>
<td>n (%)</td>
<td>n (%)</td>
</tr>
<tr>
<td>Total RIT</td>
<td>85 (39.17)</td>
<td>132 (60.83)</td>
</tr>
<tr>
<td>Lit RIT</td>
<td>87 (40.10)</td>
<td>130 (59.90)</td>
</tr>
<tr>
<td>IT RIT</td>
<td>92 (42.40)</td>
<td>125 (57.60)</td>
</tr>
<tr>
<td>VAU RIT</td>
<td>88 (40.55)</td>
<td>129 (59.45)</td>
</tr>
<tr>
<td>Spring 2017</td>
<td>n (%)</td>
<td>n (%)</td>
</tr>
<tr>
<td>Total RIT</td>
<td>80 (34.78)</td>
<td>150 (65.28)</td>
</tr>
<tr>
<td>Lit RIT</td>
<td>83 (36.09)</td>
<td>147 (63.91)</td>
</tr>
<tr>
<td>IT RIT</td>
<td>92 (40.00)</td>
<td>138 (60.00)</td>
</tr>
<tr>
<td>VAU RIT</td>
<td>81 (35.22)</td>
<td>149 (64.78)</td>
</tr>
<tr>
<td>eMPowerME Spring 2017</td>
<td>n (%)</td>
<td>n (%)</td>
</tr>
<tr>
<td>Sample</td>
<td>145 (61.70)</td>
<td>90 (38.30)</td>
</tr>
<tr>
<td>State*</td>
<td>(47.77)</td>
<td>(52.23)</td>
</tr>
</tbody>
</table>

Note: *Source 2016-2017 eMPowerME ELA/Literacy & Mathematics Technical Report. ELA/L - English Language Arts/Literacy; MAP-R = Measures of Academic Progress-Reading; RIT = Rasch Unit; Lit = literature; IT = informational text; VAU = vocabulary acquisition and usage.*
Table 7 presents the descriptive statistics (i.e., overall means and standard deviations) for MAP-R total scores and RIT subscores and the eMPowerME ELA/Literacy scaled scores, for students in my sample, students in the MAP-R national norm group, and students in the state of Maine who took eMPowerME ELA/L in spring 2017. According to the NWEA (2015) normative data, Grade 6 students in the middle of the "end-year" period had a mean reading total RIT score of 215.8 (SD = 14.66), Grade 7 students had mean RIT score of 218.2 (SD of 15.14), and Grade 8 students had mean RIT score of 220 (SD of 15.73). As can be seen in Table 7, these data indicated that students in the Grade 6-7 cohort scored slightly below and students in the Grade 7-8 cohort scored somewhat above the national norms in the spring of 2016 and 2017. When comparing the total RIT for both cohorts as seventh graders, the Grade 7-8 cohort mean score was 3.25 RIT units higher than the mean score for the Grade 6-7 cohort. Within the goal areas, both seventh-grade groups had the highest mean scores on the vocabulary acquisition and use subtest, however, the gap between mean scores was also the highest in this area.

Because statewide eMPowerME ELA/L grade-level mean scaled scores were not available for direct comparison, I used the raw-to-scaled score correspondence data available for Grades 7 and 8 in the eMPowerME technical report and converted the scaled scores in my sample to raw scores. To determine how far above or below the sample mean was relative to the state mean, I subtracted the sample raw mean from the state raw mean and divided that value by the state SD. The mean raw score for Grade 7 and 8 in the sample were 0.34 and 0.36 standard deviation units below the state raw mean for Grades 7 and 8, respectively. Again, because the individual statewide scores were not available, I was not able to determine if this difference was statistically significant.
Table 7

Descriptive Statistics for MAP-R Total and Subtests and eMPowerME-ELA/L

<table>
<thead>
<tr>
<th></th>
<th>Grade Cohort</th>
<th>6-7</th>
<th>7-8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>n</td>
<td>M</td>
</tr>
<tr>
<td>MAP-R</td>
<td></td>
<td>Spring 2016</td>
<td></td>
</tr>
<tr>
<td>Total RIT</td>
<td></td>
<td>217</td>
<td>214.74</td>
</tr>
<tr>
<td>Lit RIT</td>
<td></td>
<td>217</td>
<td>214.57</td>
</tr>
<tr>
<td>IT RIT</td>
<td></td>
<td>217</td>
<td>213.72</td>
</tr>
<tr>
<td>VAU RIT</td>
<td></td>
<td>217</td>
<td>215.89</td>
</tr>
<tr>
<td>Spring 2017</td>
<td></td>
<td>230</td>
<td>217.03</td>
</tr>
<tr>
<td>Total RIT</td>
<td></td>
<td>230</td>
<td>216.22</td>
</tr>
<tr>
<td>Lit RIT</td>
<td></td>
<td>230</td>
<td>216.57</td>
</tr>
<tr>
<td>IT RIT</td>
<td></td>
<td>230</td>
<td>218.25</td>
</tr>
<tr>
<td>VAU RIT</td>
<td></td>
<td>230</td>
<td></td>
</tr>
<tr>
<td>eMPowerME Spring 2017</td>
<td>235</td>
<td>755.68</td>
<td>14.08</td>
</tr>
<tr>
<td>Sample Scaled Score</td>
<td>235</td>
<td>29.00</td>
<td>–</td>
</tr>
<tr>
<td>Sample Raw Score</td>
<td></td>
<td>235</td>
<td>32.98</td>
</tr>
<tr>
<td>State Raw Score*</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: *Source 2016-2017 eMPowerME ELA/Literacy & Mathematics Technical Report. ELA/L - English Language Arts/Literacy; MAP-R = Measures of Academic Progress-Reading; RIT = Rasch Unit; Lit = literature; IT = informational text; VAU = vocabulary acquisition and usage.

Table 8 shows the eMPowerME ELA/L mean scores and standard deviations by student demographics for the Grade 6-7 and 7-8 cohorts. The mean score was higher for female students than for male students, higher for students who do not receive FRL than for those students who do receive FRL, and higher for students without an IEP compared to those who do have an IEP.
Table 8

**Descriptive Statistics eMPowerME Spring 2017 by Student Demographics**

<table>
<thead>
<tr>
<th>Student Characteristics</th>
<th>Grade Cohort</th>
<th>6-7</th>
<th>7-8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Male</td>
<td>123</td>
<td>753.28</td>
<td>14.58</td>
</tr>
<tr>
<td>Female</td>
<td>112</td>
<td>758.32</td>
<td>13.08</td>
</tr>
<tr>
<td>No FRL</td>
<td>122</td>
<td>759.88</td>
<td>13.92</td>
</tr>
<tr>
<td>FRL</td>
<td>113</td>
<td>751.15</td>
<td>12.85</td>
</tr>
<tr>
<td>No IEP</td>
<td>186</td>
<td>758.91</td>
<td>12.86</td>
</tr>
<tr>
<td>IEP</td>
<td>49</td>
<td>743.41</td>
<td>11.62</td>
</tr>
</tbody>
</table>

*Note. FRL = free/reduced price lunch; IEP = Individualized Education Program*

**Correlation Analysis**

I computed a Pearson product-moment correlation coefficient to measure the strength and direction of the linear relationship among the measures of reading for both cohorts to address the first and second questions. Before running the correlation analyses, I considered critical assumptions related to the study design, including the two variables must be paired and continuous as well as how the data fits the model. After visual inspection of the scatterplot, I concluded that there were positive linear relationships between both the spring 2016 and spring 2017 MAP-R and spring eMPowerME ELA/L scores. I used several pieces of evidence to determine whether or not the data were normally distributed. I assessed the skewness and kurtosis values for the MAP-R and eMPowerME ELA/L by applying the guideline for skewness value as between -1 and 1 and the kurtosis value as between -2 and 2.

Spring 2017 eMPowerME scores were normally distributed for the Grade 6-7 and 7-8 cohorts with a skewness of 0.015 (SE = 0.159) and -0.366 (SE = 0.158) and kurtosis of -0.346 (SE = 0.316) and 0.601 (SE = 0.314), respectively. The value of skewness and kurtosis were
within the normal range for both Spring 2016 and 2017 MAP-R scores as well. The Grade 6-7 cohort had a skewness of -0.718 \( (SE = 0.165) \) and -0.762 \( (SE = 0.160) \) and kurtosis of 1.201 \( (SE = 0.329) \) and 0.747 \( (SE = 0.320) \). The Grade 7-8 cohort had a skewness of -0.522 \( (SE = 0.163) \) and -0.616 \( (SE = 0.165) \) and kurtosis of 0.782 \( (SE = 0.325) \) and 0.673 \( (SE = 0.328) \).

I visually inspected the eMPowerME ELA/L and MAP-R histograms and Normal Q-Q Plots and found the scores were approximately normally distributed for both grade-level cohorts. Finally, I examined histograms and boxplots for outliers with values greater than 1.5 box-lengths from the edge of the box (Field, 2013) and found four within the eMPowerME and eight within the MAP-R data sets across both cohorts. I reviewed each score and found them to be accurate and appropriate. Furthermore, I ran a Pearson's correlation analysis with and without the mild outliers, compared the results, and determined that the outlier did not have a considerable effect on my analysis; therefore, I continued the investigation with all cases.

The purpose of running a Pearson correlation procedure was to acquire the value of the Pearson correlation coefficient. I examined the output and determined there were positive correlations between MAP-R and eMPowerME ELA/L scores for the Grade 6-7 and 7-8 cohorts. As seen in Table 8, the MAP-R scores showed strong associations with performance on the eMPowerME ELA/L, indicating reasonable concurrent validity. According to Cohen (1988), values between 0.1 and 0.3 are considered weak values between 0.3 and 0.5 are moderate, and 0.5 and greater are strong correlations. The coefficient of determination was the proportion of variance in one variable that was explained by the other variable and was calculated as the square of the correlation coefficient \( (r^2) \). For the spring 2016 MAP-R, the coefficients of determination, \( r^2 \), for the Grade 6-7 and 7-8 cohorts were equal to \( 0.75^2 = 0.56 \) and \( 0.80^2 = 0.64 \). For the spring 2017 MAP-R, the coefficient of determination, \( r^2 \), were both equal to \( 0.78^2 = 0.61 \). These values
can be expressed as a percentage. Thus, the MAP-R scores statistically explained 56%, 64%, and 61% of the eMPowerME ELA/L scores. In other words, there were statistically significant, strong positive correlations between spring 2016 MAP-R and spring 2017 eMPowerME scores for seventh- and eighth-graders at $r(215) = .75, p < .001$ and $r(220) = .80, p < .001$, respectively. Similarly, an increase in scores on the spring 2017 MAP-R was highly correlated with an increase on spring 2017 eMPowerME ELA/L scores for both cohorts, $r(228) = .78, p < .001$ and $r(216) = .78, p < .001$.

The MAP-R total RIT scores showed the strongest relationships for both cohorts. However, there was some variation in relationships across cohorts and testing year. With scores on the spring 2017 eMPowerME ELA/L, the spring 2017 MAP-R total RIT scores had the strongest relationships for the Grade 6-7 cohort, whereas spring 2016 MAP-R total RIT scores had the most robust relationship for Grade 7-8 cohort. Specifically, both cohorts had the strongest correlations between MAP-R and eMPowerME ELA/L during spring of their seventh-grade year.

As seen in Table 9, I further analyzed the strength of relationships of the subtest scores (i.e., literary text, information text, vocabulary acquisition, and usage) and concluded that sixth-grade literary text scores from the Grade 6-7 cohort had the most robust relationship to the state test performance. Informational text scores were strongest for seventh graders in both cohorts and eighth graders from the 7-8 cohort. Vocabulary acquisition and usage had the weakest relationships in both cohorts at all three grades but were still moderately strong. Overall, there were strong, positive correlations between all MAP-R scores and eMPowerME ELA/L scores. Higher scores on the MAP-R were correlated with higher scores on the eMPowerME.
Table 9

*Correlations Matrix of MAP–R and eMPowerME ELA/L*

<table>
<thead>
<tr>
<th></th>
<th>eMPowerME Spring 2017</th>
<th>Grade Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6-7</td>
<td>7-8</td>
</tr>
<tr>
<td><strong>MAP–R</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring 2016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total RIT</td>
<td>.75**</td>
<td>.80**</td>
</tr>
<tr>
<td>Lit RIT</td>
<td>73**</td>
<td>75**</td>
</tr>
<tr>
<td>IT RIT</td>
<td>.70**</td>
<td>.78**</td>
</tr>
<tr>
<td>VAU RIT</td>
<td>.68**</td>
<td>.73**</td>
</tr>
<tr>
<td>Spring 2017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total RIT</td>
<td>.78**</td>
<td>.78**</td>
</tr>
<tr>
<td>Lit RIT</td>
<td>.74**</td>
<td>.73**</td>
</tr>
<tr>
<td>IT RIT</td>
<td>.76**</td>
<td>.75**</td>
</tr>
<tr>
<td>VAU RIT</td>
<td>.71**</td>
<td>.71**</td>
</tr>
</tbody>
</table>

*Note.* ELA/L - English Language Arts/Literacy; MAP–R = Measures of Academic Progress-Reading; RIT = Rasch Unit; Lit = literature; IT = informational text; VAU = vocabulary acquisition and usage. **p < .01.

**Logistic Regression Analysis**

I conducted separate binomial logistic regression analyses to address the second and third research questions regarding the effectiveness of the MAP–R scores as an individual predictor and MAP–R scores combined with demographic predictor variables when determining sixth- and seventh-graders performance (proficient–not proficient) one year later on the state summative assessment, eMPowerME ELA/L. A binary logistic regression attempts to predict the log-odds that an observation falls into one of two categories of a dichotomous dependent variable based on one or more independent variables that can be either continuous or categorical (Agresti, 2018; Hair, Black, Babin, Anderson, & Tatham, 2006). The regression estimates the log-odds of an event occurring. Percentage accuracy in classification or classification accuracy (CA), sensitivity (SEN), specificity (SPE), positive predictive value (PPV), and negative predictive value (NPV)
are measures that assess the ability of a binomial logistic regression model to classify cases correctly. These measures were based on a cut-off point of 0.5 (50%). This process means that cases with a predicted probability of the event were greater than or equal to 0.5 were classified (by SPSS Statistic) as having the event (in this study, not meeting proficiency standards) and cases below 0.5 were classified as not having the event (meeting proficiency standards).

Before analyzing my data using binomial logistic regression, I considered several critical assumptions about the study design and nature of the data. To meet the assumption of the dependent variable, I collapsed the eMPowerME ELA/L achievement level categories into two groups. Student performance that fell within the well-below or below state expectations were considered not proficient and coded as 0 and performance that fell within the at or above state expectations were considered as proficient and coded as 1. For the independent variables, I utilized the continuous MAP-R total RIT score and the categorical variables including gender (female, coded as 0; male, coded as 1), SES (No FRL, coded as 0; FRL coded as 1), and disability status (no IEP, coded as 0; IEP, coded as 1). This study met the assumption of independence of observations. Similarly, with the dichotomous independent variables, a student could either be one or the other and could not be entered in both. For example, a student was identified as not having an IEP or having an IEP. To determine the minimum sample size, I used Green’s (1991) recommended formula (where \( m \) was the number of independent variables) for regression analysis. The sample size for the Grade 6-7 (\( n = 217 \)) and Grade 7-8 (\( n = 222 \)) satisfied the minimum requirement for this study (\( n > 108 \)).

For a binomial logistic regression to be valid, any continuous independent variable must be linearly related to the logit of the dependent variable. I tested this assumption using the Box-Tidwell (1962) procedure. To do this, I first performed a natural log transformation on the
independent continuous variable, spring 2016 MAP-R scores. Next, I created an interaction term for the continuous independent variable and its respective natural log-transformed variable. I then used the binary logistic procedure in SPSS statistic to produce the results for the Box-Tidwell (1962) test for MAP-R alone, and then the MAP-R scores combined with the demographic variables. From the output, I specifically consulted the Variables in the Equation table generated for both grade-level cohorts. For the first analysis using the single independent variable, the level of significance for the interaction term was $p = .296$ for the Grade 6-7 cohort and $p = .829$ for the Grade 7-8 cohort. I applied a Bonferroni correction using all three terms in the model— the continuous independent variable, the interaction term, and the intercept – resulting in statistical significance being accepted when $p < .016667$ (i.e., $.05 \div 3$). Based on this new level of acceptance of statistical significance, I concluded that the continuous independent variable, MAP-R, was linearly related to the logit of the dependent variable, eMPowerME ELA/L. For the second analysis with all four independent variables included, the level of significance for the interaction term was $p = .127$ for the Grade 6-7 cohort and $p = .856$ for the Grade 7-8 cohort. I used Bonferroni correction using all six terms in the model— three categorical independent variables, the continuous independent variable, the interaction term, and the intercept – resulting in statistical significance being accepted when $p < .00833$ (i.e., $.05 \div 6$). Based on this new level of acceptance of statistical significance, the MAP-R for both cohorts were linearly related to the logit of the eMPowerME ELA/L.

I checked for individual cases that did not fit the model well using casewise diagnostics in SPSS Statistics. Across both models, five cases from the Grade 6-7 and seven cases from the 7-8 cohorts had standardized residuals with values ranging from -3.468 to +5.520, which was greater than $\pm2.5$ standard deviations. With further investigation, I determined that the
demographics of the outlier group was representative of the entire sample. As such, I kept any outliers in the analysis. After checking the seven assumptions, I decided to proceed with my analyses.

The first logistic regression model, with the MAP-R scores as the singular predictor variable, was statistically significant for both cohorts in Grade 6-7, $\chi^2(1) = 111.389, p < .001$ and Grade 7-8, $\chi^2(1) = 123.542, p < .001$. The model explained 54.6% and 57.8% (NagelKerke $R^2$) of the variance in proficiency and correctly classified 82.5% and 79.7% of cases for the Grades 6-7 and 7-8, respectively. As can be seen in Table 10, the true positive rate (i.e., sensitivity) for both cohorts was at the higher end of the acceptability range of a screening measure of .70 to .90 (e.g., Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009; Glover & Albers, 2007; Kilgus et al., 2014). Some researchers have argued that high levels of sensitivity are necessary for universal screening (Compton at al., 2006; Jenkins et al., 2007). For instance, Johnson and colleagues (2009) recommended using a sensitivity level of .90 for high-stakes decisions. Sensitivity levels of .90, however, are also susceptible to other limitations, such as low specificity (Johnson et al., 2010).

The true negative rate (i.e., specificity) for the Grade 6-7 cohort was above the value generally considered adequate for screening purposes; however, for the Grade 7-8 cohort, it was slightly below the acceptability level of a minimum value of .70 (Kilgus et al., 2014). I used the Wald test to further substantiate the statistical significance for the independent (predictor) variable. As shown in Table 10, the MAP-R scores were statistically significant for both Grade 6-7 and 7-8 cohorts. The odds ratio indicated that for every point higher a student scored on the spring MAP-R as a sixth or seventh grader, the odds that he or she would achieve proficiency on the spring eMPowerME ELA/L in Grade 7 or 8 increased by 1.2 times.
The second logistic regression model with the MAP-R and three demographic variables was also statistically significant for both cohorts in Grade 6-7, \( \chi^2 (4) = 121.251 \ p < .001 \), and Grade 7-8, \( \chi^2 (4) = 136.77, \ p < .001 \). The model explained 58.2 and 62.2\% (NagelKerke R^2) of the variance in proficiency, which was overall slightly higher than the first model. The CA was similar to the first model but slightly decreased for the Grade 6-7 cohort and slightly improved for the Grade 7-8 cohort. The sensitivity and specificity both decreased for the Grade 6-7 cohort when student characteristics were included in the model but remained within the acceptable range for a universal screener as can be seen in Table 10. For the Grade 7-8 cohort, the sensitivity was the same as the first model, and specificity improved with all variables combined.

As shown in Table 10, of the four predictor variables, MAP-R and FRL were statistically significant, but gender and IEP did not add significantly to the model for the Grade 6-7 cohort. The odds of meeting proficiency on the state test was 2.4 times higher for those students who do not receive free or reduced-price lunch as opposed to those who have FRL status. For the Grade 7-8 cohort, the MAP-R and gender were statistically significant, but FRL and IEP did not add significantly to the model. The odds of meeting proficiency on the state test was 3.2 times greater for female as opposed to male students. Like the first model, the odds ratio indicated that for every point higher a student scored on the spring MAP-R in Grades 6 or 7, he or she was 1.2 times more likely to achieve proficiency standards on the spring eMPowerME ELA/L as seventh or eighth graders.
Table 10

Logistic Regression of Spring 2016 MAP-R and Spring 2017 eMPowerME ELA/L

<table>
<thead>
<tr>
<th>MAP-R</th>
<th>Single Predictor</th>
<th>B (SE)</th>
<th>Wald</th>
<th>p</th>
<th>OR</th>
<th>CA</th>
<th>SEN</th>
<th>SPE</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gr. 6-7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>.18 (.03)</td>
<td>50.00</td>
<td>&lt;.001</td>
<td>1.20</td>
<td>.83</td>
<td>.87</td>
<td>.76</td>
<td>.86</td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td>Gr. 7-8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>.19 (.03)</td>
<td>52.15</td>
<td>&lt;.001</td>
<td>1.21</td>
<td>.80</td>
<td>.88</td>
<td>.67</td>
<td>.80</td>
<td>.79</td>
</tr>
<tr>
<td>MAP-R Combined Predictors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gr. 6-7</td>
<td>Gender</td>
<td>.29 (.38)</td>
<td>.58</td>
<td>.45</td>
<td>1.43</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRL</td>
<td>.89 (.39)</td>
<td>5.12</td>
<td>.02</td>
<td>2.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IEP</td>
<td>1.44 (.89)</td>
<td>2.62</td>
<td>.11</td>
<td>4.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MAP-R</td>
<td>1.78 (.03)</td>
<td>42.74</td>
<td>&lt;.001</td>
<td>1.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.83</td>
<td>.70</td>
<td>.82</td>
<td>.72</td>
<td></td>
</tr>
<tr>
<td>Gr. 7-8</td>
<td>Gender</td>
<td>1.17 (.41)</td>
<td>8.25</td>
<td>.004</td>
<td>3.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FRL</td>
<td>.37 (.41)</td>
<td>.84</td>
<td>.36</td>
<td>.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IEP</td>
<td>.96 (.75)</td>
<td>1.63</td>
<td>.20</td>
<td>.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MAP-R</td>
<td>.91 (.03)</td>
<td>45.76</td>
<td>&lt;.001</td>
<td>1.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.88</td>
<td>.76</td>
<td>.85</td>
<td>.81</td>
<td></td>
</tr>
</tbody>
</table>

Note: Gender is male compared to females; FRL is Yes FRL compared to No FRL; IEP is Yes IEP compared to No IEP. OR = odds ratio; CA = classification accuracy; SEN = sensitivity; SPE = specificity; PPV = positive predictive value; NPV = negative predictive value

ROC Curve Analysis

To address question four, I tested the validity of the scores on the MAP-R universal screening measure in the prediction of eMPowerME ELA/L risk by generating receiver operating characteristic (ROC) curves. To generate the ROC curve analysis, I entered the continuous independent variable, spring 2016 NWEA MAP-R total RIT scores, and the dichotomous dependent variable, spring 2017 eMPowerME ELA/L total score (i.e., where 0 = Not Proficient and 1 = Proficient) into SPSS Statistics. The ROC curves were a measure of the overall discriminatory ability of the binomial logistic regression. Specifically, ROC curves indicated the
The probability that a student was correctly classified as at some level of risk on both eMPowerME ELA/L and the screening measure (i.e., sensitivity), and the probability that a student was correctly classified as not at some level of risk on the eMPowerME ELA/L and the screening measure (i.e., specificity). The binomial regression analysis used a single cut-off point of 0.5 (50%), whereas ROC analysis was based on all possible cut-off points in the data.

Figure 1 shows a visual representation of the ROC curve, which entails plotting the true-positive rate (sensitivity) on the y-axis and the false-positive rate (1-specificity) on the x-axis for each possible screening cut score (Hilbe, 2009). Each cut-off point changes the sensitivity and specificity of the test. For example, a higher cut-off point will increase specificity but decrease sensitivity making it harder to classify as not meeting proficiency, but easier to classify as proficient. In addition to diagnostic accuracy statistics, the area under the curve (AUC) was an overall measure of the accuracy of the predictor. AUC is a probability index that can range from 0.5 (no diagnostic ability) to 1.0 (perfect diagnostic ability). According to Hosmer and colleagues (2013), an AUC value between 0.50 and 0.70 has poor discrimination, between 0.70 and 0.80 is considered acceptable, between 0.80 and 0.90 is excellent, and above 0.90 is regarded as outstanding discrimination. The AUC for the Grade 6-7 cohort was .88, 95% CI [.838 to .928], which is an excellent level of discrimination (Hosmer et al., 2013). Also considered an excellent level of consideration, the AUC was .89, 95% CI [.854 to .934] for the Grade 7-8 cohort.

As I mentioned previously, various guidelines exist for acceptable levels of diagnostic accuracy (e.g., Compton et al., 2006; Jenkins et al., 2007; Johnson et al., 2010). In a recent meta-analysis, Kilgus et al. (2014) suggested a .80 for sensitivity and .70 for specificity when used for
Figure 1. Roc Curves Spring 2016 MAP-R and Spring 2017 eMPowerME ELA/L
screening decisions, while others like Johnson and colleagues (2009) recommended using a sensitivity level of .90 for high-stakes decisions. For example, based on a .90 sensitivity, 90% of students who did not meet proficiency on the state test would be correctly identified, but 10% who were not proficient would be misidentified as proficient. A .70 specificity would correctly identify 70% of the students who did meet proficiency, but 30% of students who were proficient on the state test would be misidentified as not proficient.

Table 11

**MAP-R Cut Scores, Sensitivity, and Specificity Associated with Percentiles**

<table>
<thead>
<tr>
<th>Spring 16 MAP-R</th>
<th>Cut Score</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Optimal (80th Percentile)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 6-7</td>
<td>224.5</td>
<td>.93</td>
<td>.51</td>
</tr>
<tr>
<td>Grade 7-8</td>
<td>227.5</td>
<td>.91</td>
<td>.64</td>
</tr>
<tr>
<td><strong>Maximized (60th Percentile)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 6-7</td>
<td>219.5</td>
<td>.84</td>
<td>.80</td>
</tr>
<tr>
<td>Grade 7-8</td>
<td>222.5</td>
<td>.81</td>
<td>.82</td>
</tr>
<tr>
<td><strong>40th Percentile</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 6-7</td>
<td>212.5</td>
<td>.29</td>
<td>.99</td>
</tr>
<tr>
<td>Grade 7-8</td>
<td>214.5</td>
<td>.22</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>20th Percentile</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 6-7</td>
<td>203.5</td>
<td>.29</td>
<td>.99</td>
</tr>
<tr>
<td>Grade 7-8</td>
<td>205.5</td>
<td>.22</td>
<td>1.0</td>
</tr>
</tbody>
</table>

As seen in Table 11, I identified the cut scores with at least .90 sensitivity (optimal) and cut scores balanced between sensitivity and specificity (maximized), with a minimum specificity of .70, for the MAP-R spring 2016 administration of both cohorts. With sensitivity above .90, specificity was negatively impacted by increasing the number of false positives (i.e., indicated
students at risk when they were not). Maximized sensitivity and specificity fell within the acceptable range (ranging from .80 to .84) for a universal screener (Kilgus et al., 2014). I compared the generated MAP-R cut scores from this sample to the MAP-R published national norms (NWEA, 2015). The optimal cut scores (.90 sensitivity) were associated with the 80th percentile (distinguishing between high-average and high) and maximized sensitivity and specificity were associated with 60th percentile (distinguishing between average and high-average). I also compared the scores from the established MAP-R norms at the 20th percentile (distinguishing low from low-average) and at the 40th percentile (distinguishing low-average from average) to the ROC curves generated from my sample. The cut scores at these ranges resulted in high specificity but poor sensitivity, indicating unacceptable rates of false negatives (i.e., failing to identify the students truly at risk).

**ANOVA and t-Test Analysis**

I used a one-way repeated measures analysis of variance (ANOVA) and independent-samples t-tests to answer the fifth question as to whether or not any statistically significant differences existed between student performance on the three MAP-R subtests and, if so, whether the differences varied by student characteristics. To run the analyses, I considered the assumptions related to the study design. For the ANOVA, the dependent variable (MAP-R subtest scores), was measured at the continuous level and the within-subjects factor consisted of three categorical levels; in this case, the three MAP-R subtests. I ran 18 separate t-tests with each of the MAP-R subtest scores as the dependent variable and each student characteristic as the independent variable using spring 2016 and spring 2017 data. I discussed the assumptions related to how the data fit the model, including no significant outliers, normal distribution, sphericity,
and homogeneity of variances. I present the results of the ANOVA and t-tests analyses in the next sections.

**Differences Between MAP-R Subtests**

I assessed outliers in the data via inspection of a boxplot for values greater than 1.5 box-lengths from the edge of the box and found one extreme case from the Grade 7-8 cohort. I ran a one-way repeated measure ANOVA with and without the extreme outlier, compared the results, and determined that the outlier did not have an appreciable effect on my analysis and thus, continued with the analysis with all cases. I visually inspected the histograms of the MAP-R subtests and concluded that the scores were normally distributed for both cohorts. The assumption of sphericity was that the difference between the levels of the within-subject factor (i.e., MAP-R scores) have equal variances. Mauchly’s test of sphericity indicated that the assumption of sphericity was not violated for either the Grade 6-7 cohort, $\chi^2(2) = 3.308, p = .191$ or the Grade 7-8 cohort, $\chi^2(2) = 2.043, p = .301$.

**MAP-R Spring 2016.** The results of the ANOVA indicated the MAP-R mean scores were statistically different for the three reading measures, $F(2, 432) = 6.407, p = .002$, partial $\eta^2 = .029$ and $F(2, 442) = 8.099, p < .001$, $\eta^2 = .035$, for the Grade 6-7 and 7-8 cohorts, respectively. Specifically, the mean of at least one of the subtests was different from the mean of at least one other subtest. A Bonferroni post hoc test is recommended (Maxwell & Delaney, 2004) for testing all pairwise combinations of levels of the within-subjects factor. This test was useful because it provided both the statistical significance level for each pairwise comparison as well as the confidence intervals for the mean difference for each comparison. For the sixth graders, the Bonferroni post hoc test results indicated that the informational text scores ($M = 213.72, SD = 14.55$) were lower than the vocabulary acquisition and use scores ($M = 215.89, SD = 15.67$) with
a statistically significant mean difference of 2.17, 95% CI [3.72, 6.07], \( p = .003 \). The seventh graders had two pairwise combinations that were statistically different. The Bonferroni post hoc test results indicated that the literary text scores (\( M = 219.96, SD = 14.78 \)) and informational text scores (\( M = 219.31, SD = 14.33 \)) were lower than the vocabulary acquisition and use scores (\( M = 221.63, SD = 15.05 \)) with a statistically significant mean difference of 1.66 (95% CI [0.19, 3.14], \( p = .021 \)) and 2.32 (95% CI [0.86, 3.78], \( p < .001 \)) respectively. There was no statistically significant difference in means scores between literary text and vocabulary acquisition and usage for the sixth graders who took the spring 2016 MAP-R. Additionally, there were no statistically significant differences in means scores between spring 2016 MAP-R literary text and informational text for either sixth or seventh graders.

**MAP-R Spring 2017.** After inspection of the data via boxplots and histograms, I determined that there were no extreme outliers in either grade cohort and both had normal distributions. Mauchly’s test of sphericity indicated that the assumption of sphericity was not violated for either seventh graders, \( \chi^2(2) = 1.183, p = .554 \) or eighth graders, \( \chi^2(2) = 1.511, p = .470 \). The results of the ANOVA indicated the MAP-R mean scores were statistically different for the three reading measures, \( F(2, 458) = 6.166, p = .002 \), partial \( \eta^2 = .026 \) and \( F(2, 434) = 4.435, p = .012 \), partial \( \eta^2 = .020 \), for seventh and eighth graders, respectively.

The seventh graders had two pairwise combinations that were statistically different. The Bonferroni post hoc test results indicated that the literary text scores (\( M = 216.22, SD = 17.35 \)) and informational text scores (\( M = 216.57, SD = 17.10 \)) were lower than the vocabulary acquisition and use scores (\( M = 218.25, SD = 16.49 \)) with a statistically significant mean difference of 2.03 (95% CI [0.49, 3.57], \( p = .005 \)) and 1.68 (95% CI [0.21, 3.12], \( p = .019 \)) respectively. The eighth graders also had two pairwise combinations that were statistically
different. The Bonferroni post hoc test results indicated that the literary text scores ($M = 221.44$, $SD = 15.49$) and informational text scores ($M = 221.53$, $SD = 14.78$) were lower than the vocabulary acquisition and use scores ($M = 223.14$, $SD = 14.71$) with a statistically significant mean difference of 1.70 (95% CI [0.11, 3.30], $p = .032$) and 1.62 (95% CI [0.39, 3.19], $p = .043$). There were no statistically significant differences between literary text and informational text scores on the spring 2017 MAP-R for either seventh or eighth graders.

**Differences within MAP-R Subtests by Student Demographics**

After testing the assumptions of the data – no significant outliers in the data, as assessed by inspection of a boxplot, and MAP-R scores were normally distributed, as assessed by visual inspection of histograms – I ran separate independent-samples t-tests to determine if there were any significant differences in spring 2016 or spring 2017 MAP-R subtest scores by student demographics. In some cases, the assumption of homogeneity of variances was violated, indicating that the population variance for each group of the independent variable was different (Laerd Statistics, 2015). SPSS not only provides an independent t-test that is calculated normally (with pooled variances) but also another t-test when the assumption is violated that uses non-pooled variances. As such, I used the appropriate t-test provided depending on whether or not the assumption of homogeneity of variances was violated. Finally, the independent t-test informs whether a difference between two groups is statistically significant, but it does not determine the size of the difference. To overcome this limitation, I determined the effect size with all t-tests that had statistically significant results. An effect size is the measure of the strength of a relationship between two variables (Cohen, 1988). I used Cohen’s $d$ by hand calculating the difference between the means divided by the pooled SD. According to Cohen (1988), a $d$ of 0.2 is small, 0.5 is medium, and 0.8 is a large effect size.
**Literary Text Comprehension – Gender.** Females scored higher than males on the MAP-R literary text subtest for Grades 6, 7 and 8 with three out of the four sets of data that I compared showing statistically significant differences. There was a statistically significant difference in spring 2016 MAP-R literary text scores \( (M = 8.35, 95\% \text{ CI} [4.28 \text{ to } 12.42], t(215) = 4.05 \ p < .001, \ d = .55) \) between sixth-grade females \( (M = 219.07 \ SD = 11.71, \ n = 100) \) and males \( (M = 210.72, \ SD = 17.58, \ n = 117) \). There was also a statistically significant difference in spring 2017 MAP-R literary text scores \( (M = 6.01, 95\% \text{ CI} [1.56 \text{ to } 10.47], t(228) = 2.66 \ p < .008, \ d = .35; \ M = 6.32, 95\% \text{ CI} [2.26 \text{ to } 10.38], t(216) = 3.07 \ p < .002, \ d = .42) \), between seventh-grade females \( (M = 219.39, \ SD = 14.65, \ n = 109) \) and males \( (M = 213.37, \ SD = 19.07, \ n = 121) \), and eighth-grade females \( (M = 224.69, \ SD = 14.38, \ n = 106) \) and males \( (M = 218.37, \ SD = 15.94, \ n = 112) \). There was not, however, a statistically significant difference in spring 2016 MAP-R literary text scores \( (M = .96, 95\% \text{ CI} [3.81 \text{ to } 4.01], t(219.97) = .05, \ p = .958) \) between seventh-grade females \( (M = 220.02, \ SD = 14.19, \ n = 106) \) and males \( (M = 219.91, \ SD = 15.36, \ n = 116) \).

**Literary Text Comprehension – Socioeconomic Status (SES).** When comparing performance by SES, students who came from economically advantaged backgrounds consistently scored higher than their peers who came from economically disadvantaged backgrounds on the MAP-R literary text subtest. For sixth graders, there was a statistically significant difference in spring 2016 MAP-R literary text scores \( (M = 7.83, 95\% \text{ CI} [3.73 \text{ to } 11.92], t(215) = 3.76, \ p < .001, \ d = .51) \) between students without FRL status \( (M = 218.10, \ SD = 14.05, \ n = 119) \) and those who did receive FRL \( (M = 210.28, \ SD = 16.56, \ n = 98) \). Seventh graders without FRL status \( (M = 222.94, \ SD = 14.20, \ n = 120; \ M = 220.83, \ SD = 15.41, \ n = 122) \) and with FRL status \( (M = 216.46, \ SD = 14.74, \ n = 102; \ M = 211.02, \ SD = 17.10, \ n = 108) \) were statistically significant in spring 2016 \( (M = 6.48, 95\% \text{ CI} [2.65 \text{ to } 10.32], t(220) = 3.33, \ p < .001, \ d = .45) \) and spring 2017
\( M = 9.81, \text{ 95\% CI [5.47 to 14.15], } t(212) = 4.41, p < .001, d = .59 \). Finally, eighth graders who did not receive FRL in spring 2017 \( M = 224.88, SD = 14.09, n = 113 \) scored statistically significantly higher on the literary text subtest \( M = 7.13, \text{ 95\% CI [2.05 to 11.17], } t(216) = 3.48, p = .001, d = .47 \) than their peers who did receive FRL \( M = 217.74, SD = 16.14, n = 105 \).

**Literary Text Comprehension – Disability Status.** Students who did not have an identified disability scored significantly higher than those who did have a disability on the MAP-R literary text subtest. When comparing spring 2016 MAP-R scores of sixth graders \( M = 19.327, \text{ 95\% CI [13.62 to 25.04] } t[53.054] = 6.79, p < .001, d = 1.26 \) and seventh graders \( M = 16.25, \text{ 95\% CI [11.79 to 20/71] } t[220] = 7.12, p < .001, d = 1.17 \) those who did not have an IEP \( M = 218.40, SD = 12.56, n = 174 \); \( M = 223.11, SD = 12.97, n = 179 \) scored higher than those who did have an IEP \( M = 199.07, SD = 17.59, n = 43 \); \( M = 206.86, SD = 14.76, n = 43 \). Likewise, the spring 2017 MAP-R scores revealed seventh graders \( M = 19.05, \text{ 95\% CI [12.62 to 25.49], } t[56.929] = 5.93, p < .001, d = 1.07 \) and eighth graders \( M = 16.71, \text{ 95\% CI [12.01 to 21.42], } t[216] = 7.01, p < .001, d = 1.18 \) without disabilities \( M = 220.11, SD = 13.93, n = 183 \); \( M = 224.74, SD = 13.91, n = 175 \) performed better than their peers with disabilities \( M = 201.06, SD = 20.87, n = 47 \); \( M = 208.02, SD = 14.45, n = 43 \) on the literary text subtest.

**Informational Text Comprehension – Gender.** Females scored higher than males on the MAP-R informational text subtest with the exception of spring 2016 administration when seventh-grade males scored slightly higher than seventh-grade females. There was a statistically significant difference in sixth graders spring 2016 scores \( M = 6.49, \text{ 95\% CI [2.76 to 10.21], } t[209.003] = 3.34, p = .001, d = .46 \) as well as eighth graders spring 2017 scores \( M = 4.01, \text{ 95\% CI [0.08 to 7.93], } t[216] = 2.01, p = .045, d = .27 \) with females \( M = 217.22, SD = 11.58, n = 100 \); \( M = 223.58, SD = 15.42, n = 106 \) scoring higher than males \( M = 210.74, SD = 16.12, n = 117 \).
There was not, however, any statistically significant differences in mean informational text scores for seventh graders in either spring 2016 ($M = .83$, 95% CI [-2.93 to 4.63], $t(220) = .43$, $p = .668$) or spring 2017; $M = 4.34$, 95% CI [.04 to 8.72] $t(225.69) = 1.96$, $p = .052$) between females ($M = 218.88$, $SD = 15.49$, $n = 106$; $M = 218.85$, $SD = 15.08$, $n = 109$) and males ($M = 219.71$, $SD = 13.23$, $n = 166$; $M = 214.51$, $SD = 18.55$, $n = 121$).

**Informational Text Comprehension – Socioeconomic Status (SES).** When comparing performance by SES, students who came from economically advantaged backgrounds consistently scored higher than their peers who came from economically disadvantaged backgrounds on the MAP-R informational text subtest. For sixth graders, there was a statistically significant difference in spring 2016 MAP-R informational text scores ($M = 5.51$, 95% CI [1.65 to 9.36] $t([215] = 2.82$, $p = .005$, $d = .38$) between students without FRL status ($M = 216.21$, $SD = 13.22$, $n = 119$) and with FRL status ($M = 210.70$, $SD = 15.56$, $n = 98$). Seventh graders who did not qualify for FRL ($M = 223.35$, $SD = 12.42$, $n = 120$; $M = 221.34$, $SD = 15.33$, $n = 122$) and who did qualify for FRL ($M = 214.56$, $SD = 15.01$, $n = 102$; $M = 211.18$, $SD = 17.45$, $n = 108$) in spring of 2016 ($M = 8.79$, 95% CI [5.16 to 12.42] $t(220) = 4.78$, $p < .001$, $d = .64$) and spring 2017 ($M = 10.17$, 95% CI [5.91 to 14.43] $t(228) = 4.70$, $p < .001$, $d = .62$) were statistically significant. Finally, eighth graders who did not receive FRL ($M = 225.21$, $SD = 12.83$, $n = 113$) in spring 2017 scored statistically significantly higher on the informational text subtest ($M = 7.65$, 95% CI [3.83 to 11.47] $t(216) = 3.95$, $p < .001$, $d = .53$) than their peers who did receive FRL ($M = 217.56$, $SD = 15.74$, $n = 105$).

**Informational Text Comprehension – Disability Status.** Students who did not have an identified disability scored significantly higher than those who did have a disability on the MAP-
R literary text subtest. When comparing spring 2016 MAP-R scores of sixth-graders ($M = 15.72$, 95% CI [10.02 to 21.43] $t[51.567] = 5.53, p < .001, d = 1.05$) and seventh-graders ($M = 16.68$, 95% CI [11.85 to 21.51] $t[220] = 7.71, p < .001, d = 1.24$) who did not have an IEP ($M = 216.84$, $SD = 11.79, n = 174; M = 222.54, SD = 12.25, n = 179$) scored higher than those who did have an IEP ($M = 210.12, SD = 17.70, n = 43; M = 205.86, SD = 14.63, n = 43$). Likewise, the spring 2017 MAP-R scores revealed that seventh-graders ($M = 19.94$, 95% CI [14.22 to 25.67] $t[61.06] = 6.97, p < .001, d = 1.22$) and eighth-graders ($M = 16.39$, 95% CI [11.93 to 20.95] $t[216] = 7.25, p < .001, d = 1.18$) without disabilities ($M = 220.64, SD = 14.21, n = 183; M = 224.76, SD = 12.92, n = 175$) performed better than their peers with disabilities ($M = 200.70, SD = 18.25, n = 47; M = 208.37, SD = 14.73, n = 43$) on the informational text subtest.

**Vocabulary Acquisition and Use – Gender.** Females scored higher than males on the MAP-R vocabulary acquisition and use subtest with the exception of spring 2017 administration when males scored slightly higher than females. There was not a statistically significant difference in scores ($M = 4.06$, 95% CI [.03 to 8.15] $t[211.76] = 1.96, p = .052$) between sixth-grade females ($M = 218.08, SD = 13.11, n = 100$) and males ($M = 214.02, SD = 17.40, n = 117$). In spring 2016 and spring 2017, there also was not a statistically significant difference between scores ($M = .15$, 95% CI [-3.84 to 4.15] $t[220] = .08, p = .941; M = 3.11$, 95% CI [1.17 to 7.40] $t[228] = 1.43, p = .153$) of seventh-grade females ($M = 221.55, SD = 15.20, n = 106; M = 219.89, SD = 13.47, n = 109$) and males ($M = 221.70, SD = 14.92, n = 116; M = 216.78, SD = 18.73, n = 121$). Again, there was not a significant difference in scores ($M = 1.08$, 95% CI [2.85 to 5.02] $t[216] = .54, p = .588$) between eighth-grade females ($M = 223.70, SD = 14.67, n = 106$) and males ($M = 222.62, SD = 14.79, n = 112$) on the vocabulary acquisition and use subtest.
Vocabulary Acquisition and Use – Socioeconomic Status (SES). When comparing performance by SES, students who came from economically advantaged backgrounds consistently scored higher than their peers who came from economically disadvantaged backgrounds on the MAP-R vocabulary acquisition and use subtest. For sixth graders, there was a statistically significant difference in spring 2016 scores ($M = 7.76, 95\% \text{ CI} [3.67 \text{ to } 11.85]$ $t[215] = 3.74, p < .001, d = .51$) between students without FRL status ($M = 219.39, SD = 14.62, n = 119$) and with FRL status ($M = 211.63, SD = 15.91, n = 98$). Seventh graders who did not qualify for FRL ($M = 225.42, SD = 14.08, n = 120$; $M = 222.89, SD = 14.88, n = 122$) and who did qualify for FRL ($M = 217.17, SD = 14.10, n = 102$; $M = 213.02, SD = 16.70, n = 108$) in spring of 2016 ($M = 8.25, 95\% \text{ CI} [4.40 \text{ to } 12.10]$ $t[220] = 4.22, p < .001, d = .57$) and spring 2017 ($M = 9.87, 95\% \text{ CI} [5.76 \text{ to } 13.97]$ $t[228] = 4.74, p < .001, d = .62$) were statistically significant. Finally, eighth graders who did not receive FRL ($M = 226.94, SD = 12.67, n = 113$) in spring 2017 scored statistically significantly higher on the vocabulary acquisition and use subtest ($M = 7.88, 95\% \text{ CI} [4.05 \text{ to } 11.70]$ $t[200.039] = 4.06, p < .001, d = .55$) than their peers who did receive FRL ($M = 219.06, SD = 15.69, n = 105$).

Vocabulary Acquisition and Use – Disability Status. Students who did not have an identified disability scored significantly higher than those who did have a disability on the MAP-R literary text subtest. When comparing spring 2016 MAP-R scores of sixth-graders ($M = 18.63, 95\% \text{ CI} [12.94 \text{ to } 24.31]$ $t[53.593] = 6.60, p < .001, d = 1.22$) and seventh-graders ($M = 17.59, 95\% \text{ CI} [13.12 \text{ to } 22.07]$ $t[220] = 7.75, p < .001, d = 1.27$) students who did not have an IEP ($M = 219.58, SD = 12.77, n = 174$; $M = 225.03, SD = 13.09, n = 179$) scored higher than those who did have an IEP ($M = 200.95, SD = 17.48, n = 43$; $M = 207.44, SD = 14.48, n = 43$). Likewise, the spring 2017 MAP-R scores revealed that seventh-graders ($M = 19.04, 95\% \text{ CI} [13.08 \text{ to } 25.00]$ $t[228] = 4.74, p < .001, d = .62$) scored significantly higher than their peers who did have an IEP ($M = 200.03, SD = 14.48, n = 43$; $M = 207.44, SD = 14.48, n = 43$).
24.99) \( t[57.480] = 6.40, p < .001, d = 1.15 \) and eighth-graders (\( M = 15.97, 95\% \text{ CI} [11.51 \text{ to } 20.43] \) \( t[216] = 7.06, p < .001, d = 1.18 \) without disabilities (\( M = 222.14, SD = 13.17, n = 183; M = 226.29, SD = 13.15, n = 175 \) performed better than their peers with disabilities (\( M = 203.11, SD = 19.28, n = 47; M = 210.33, SD = 13.88, n = 43 \) on the vocabulary acquisition and use subtest.

Hierarchical Regression Analysis

For the final research question, I conducted hierarchical regression analyses to examine the unique contribution of spring 2016 MAP-R goal areas to the spring 2017 eMPowerME ELA/L for each grade-level cohort. I examined the strength of the relationships between the continuous independent variables for multicollinearity using correlation analysis, the value of tolerance, variance inflation factor (VIF), and condition index (CI) tests in SPSS statistics. Multicollinearity is when the independent/predictor variables are highly correlated (Hair et al., 2014). In a regression analysis, the presence of multicollinearity indicates a redundant use of information in the model, which can lead to unstable regression coefficient estimates (Hair et al.). The MAP-R subtests were highly related to each other with correlations ranging from \( r = .80 \) to \( .85, p < .01 \). Correlations exceeding \( .90 \) are considered a serious multicollinearity threat (Hair et al.). Tolerance and VIF tests also can indicate a collinearity problem when the values of tolerance are less than 0.1, and the values of VIF are above 10 (Hair et al.). The tolerance values for the MAP-R subtests ranged from \( .22 \) to \( .28 \), and VIF values ranged from 3.58 to 4.65. Finally, the CIs for the MAP-R subtests ranged from 34 to 75. CIs greater than 30 exemplify a severe problem of multicollinearity among the continuous independent variables (Hair et al.). Given the high correlation and CI values, I determined that multicollinearity existed between the MAP-R subtests. Because the MAP-R subscale scores were derived from the total score, it
makes sense that the scores are related to each other, causing this issue of multicollinearity. Due to the exploratory nature of question 6, though, I determined the potential contribution of the findings to outweigh the limitations of the data. Therefore, I continued with my planned analyses.

Table 12

Summary of Hierarchical Regression Analyses Using Spring 2016 MAP-R Goal Areas to Predict Performance on Spring 2017 eMPowerME ELA/L

<table>
<thead>
<tr>
<th>Grade Cohort</th>
<th>Model</th>
<th>Step</th>
<th>Predictor</th>
<th>R^2</th>
<th>ΔR^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-7</td>
<td>A</td>
<td>1</td>
<td>Literary Text</td>
<td>.53</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Informational Text</td>
<td>.56</td>
<td>.03**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Vocabulary Acquisition and Use</td>
<td>.56</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>1</td>
<td>Informational Text</td>
<td>.49</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Literary Text</td>
<td>.56</td>
<td>.07**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Vocabulary Acquisition and Use</td>
<td>.56</td>
<td>.00</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>1</td>
<td>Vocabulary Acquisition and Use</td>
<td>.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Literary Text</td>
<td>.55</td>
<td>.08**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Informational Text</td>
<td>.56</td>
<td>.02*</td>
</tr>
<tr>
<td>7-8</td>
<td>A</td>
<td>1</td>
<td>Literary Text</td>
<td>.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Informational Text</td>
<td>.64</td>
<td>.08**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Vocabulary Acquisition and Use</td>
<td>.65</td>
<td>.01*</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>1</td>
<td>Informational Text</td>
<td>.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Literary Text</td>
<td>.64</td>
<td>.04**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Vocabulary Acquisition and Use</td>
<td>.65</td>
<td>.01*</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>1</td>
<td>Vocabulary Acquisition and Use</td>
<td>.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>Literary Text</td>
<td>.61</td>
<td>.07**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>Informational Text</td>
<td>.65</td>
<td>.04**</td>
</tr>
</tbody>
</table>

Note. *p < .01, **p < .001
I ran three separate regression models with three blocks, each rotating the subtest scores through the blocks. As seen in Table 12, the combined MAP-R subtests explained 56% and 65% of the variance in the eMPowerME ELA/L scores for the Grade 6-7 and 7-8 cohorts respectively (p < .001). For the Grade 6-7 cohort, literary text scores accounted for greatest variance (53%) in students' eMPowerME ELA/L scores when entered first, and informational text accounted for an additional 3% variance (p < .001). VAU failed to explain any significant unique variance. Informational text accounted for the greatest variance (60%) when entered first for the Grade 7-8 cohort, literary text accounted for an additional 4% variance (p < .001), and vocabulary acquisition and use added 1% (p < .01).

Summary

In this chapter, I described the results of the statistical analyses used to answer the research questions that guided this study. I computed a Pearson product-moment correlation coefficient to assess the relationship between MAP-R total and subtest scores and eMPowerME ELA/L scores. There were statistically significant, strong positive correlations between spring 2016 and spring 2017 MAP-R total RIT scores and spring 2017 eMPowerME scores (ranging from .75 to .80) and strong correlations (ranging from .68 to .78) between all MAP-R subtest scores for both grade-level cohorts. Overall, the results indicated a robust concurrent validity between MAP-R and eMPowerME ELA/L.

To further understand the predictive nature of MAP-R scores and demographic variables in determining student proficiency on the eMPowerME ELA/L, I conducted binomial logistic regression and ROC curve analyses. The first logistic regression model, with the MAP-R scores as the single predictor variable, and the second model, with the MAP-R, gender, SES, and disability status as predictors variables, were both statistically significant. The second model
explained a slightly higher variance in proficiency than the first model (variances ranging from 55% to 62%). The overall percentage accuracy in classification for both models ranged from 78.0% to 83.3%, sensitivity ranged from .83 to .88, and specificity ranged from .67 to .76. The Grade 7-8 cohort with MAP-R scores and demographic variables was the overall best-fitting model. Of the four predictor variables that resulted in statistical significance, the MAP-R scores and SES were for the Grade 6-7 cohort, and MAP-R scores and gender were for the Grade 7-8 cohort.

The odds ratio indicated that for every point higher a student scored on the spring MAP-R as a sixth or seventh grader, the odds were 1.2 times greater he or she would achieve proficiency on the spring eMPowerME ELA/L in Grade 7 or 8. The odds of meeting proficiency standards in seventh grade were 2.4 times higher for sixth graders from economically advantaged backgrounds than those who were economically disadvantaged. Seventh-grade females had 3.2 higher odds than their male counterparts at achieving proficiency standards on the eMPowerME in eighth grade.

Based on the recommendations by Kilgus et al. (2014) on the acceptability levels of a universal screening measure, the cut scores for risk that maximized sensitivity and specificity (i.e., true positives and true negatives) on MAP-R for Grades 6 and 7 were associated with the NWEA’s national norms at the 60th percentile. The cut scores that equated to the 80th percentile maximized sensitivity but decreased the specificity below the acceptable threshold. The cut scores that compared with the 40th and 20th percentiles increased the specificity well-above the acceptable level but reduced the sensitivity well below the appropriate level for a universal screening tool.
When examining sixth-, seventh-, and eighth-grade student performance on the MAP-R within and across subtests and subtest scores concerning unique variance on the eMPowerME ELA/L, several findings emerged. Firstly, there were no significant differences between literary and informational text scores. Student in Grade 6, 7, and 8 scored significantly higher on the vocabulary acquisition and use subtest compared to the informational text subtest and seventh- and eighth-graders scored significantly higher on the vocabulary acquisition as compared to the literary text subtest. Secondly, females scored higher than males on the literary text subtest with three out of the four analyses being statistically significant and having small to moderate effect sizes. Females also scored higher than males on the informational text subtest except for seventh graders in spring 2016. The gap in performance between genders was narrower when comparing performance on informational text to literary text with statistically significant differences occurring in only two out of the four analyses and producing smaller effect sizes. Again, females scored higher than males on the vocabulary acquisition and use subtest except for seventh graders in spring 2017. There was not, however, any statistically significant differences in performance between females and males on vocabulary acquisition and use subtest.

When considering SES, all subtest analyses showed statistically significant differences with large to moderate effect sizes between the performance of those who do and do not come from disadvantaged backgrounds. Differences between those with and without disabilities also were significant with large effect sizes for all subtest analyses. That is, students from disadvantaged backgrounds and with identified disabilities overall performed significantly lower on the MAP-R than their peers from advantaged backgrounds and without disabilities regardless of text type or vocabulary skills. Finally, MAP-R literary text subtest scores explained the greatest variance in eMPowerME ELA/L scores for the Grade 6-7 cohort and informational text
subscores accounted for the highest variance for the Grade 7-8 cohort. Of the three subtests, vocabulary acquisition and use had the least amount of variance for both cohorts and failed to explain significant unique variance for the Grade 6-7 cohort. Table 13 provides a summary of the research questions, data analyses, and results. In Chapter 5, I review the data analyses and results and discuss the implications of the key findings in this study.
<table>
<thead>
<tr>
<th>Questions</th>
<th>Analyses</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1.</strong> What is the concurrent validity of the MAP-R and the eMPowerME ELA/L in Grades 7 and 8?</td>
<td>Pearson’s Correlation</td>
<td>Strong correlations (.75 to .80, <em>p</em> &lt; .001) between MAP-R and eMPowerME ELA/L scores</td>
</tr>
<tr>
<td><strong>2.</strong> How useful are MAP-R scores in identifying sixth and seventh graders who will and will not meet proficiency standards on eMPowerME ELA/L one year later?</td>
<td>Pearson’s Correlation, Logistic Regression</td>
<td>Regression models were statistically significant; MAP-R explained 55% (Grades 6-7) and 58% (Grades 7-8) of the variance in proficiency and correctly classified 83% and 80% of cases on eMPowerME ELA/L; sensitivity = .87 and .88 and specificity = .76 and .67. The odds a student met proficiency on the eMPowerME ELA/L in Grade 7 or 8 increased by 1.2 times for every point higher he or she scored on the MAP-R (Grades 6 or 7).</td>
</tr>
<tr>
<td><strong>3.</strong> Is the predictive validity of the MAP-R scores improved with the inclusion of demographic variables (i.e., gender, SES status, disability status)?</td>
<td>Logistic Regression</td>
<td>Regression models were statistically significant (<em>p</em> &lt; .001); MAP-R and FRL added significantly for Grade 6-7; MAP-R and gender added significantly for Grade 7-8; MAP-R with demographic variables explained 58% (Grades 6-7) and 62% (Grades 7-8) of the variance in proficiency and correctly classified 78% and 83% of cases on eMPowerME ELA/L; sensitivity (.78 and .83) and specificity (.83 and .88). The odds of meeting proficiency on the state test in Grade 7 was 2.4 times higher for those students who do not receive free or reduced-price lunch as opposed to those who have FRL status; the odds of meeting proficiency on the state test in Grade 8 was 3.2 times greater for female as opposed to male students.</td>
</tr>
<tr>
<td><strong>4.</strong> What are the optimal cut scores for MAP-R in spring of Grades 6 and 7 for predicting performance on the eMPowerME ELA/L in Grades 7 and 8?</td>
<td>ROC Curve</td>
<td>The AUC was an excellent level of discrimination (.88 and .89). Cut scores associated with NWEA’s established risk for Grades 6 and 7 (&lt; 40th percentile) resulted in unacceptable rates of false negatives. Statistically derived cut scores (Grade 6 – 220 RIT and Grade 7 – 223 RIT) improved the diagnostic accuracy by maximizing the sensitivity/specificity; cut scores corresponded with NWEA’s established threshold at the 60th percentile.</td>
</tr>
</tbody>
</table>
Table 13 Continued.

| 5. | Are there differences between and among subtest scores on the MAP-R for sixth, seventh, and eighth graders, and if so, do the differences vary according to student demographics (i.e., gender, SES status, disability status)? | ANOVA t-Tests | Students scored higher on the MAP-R vocabulary acquisition and use subtest compared to either literary text or informational text with three out of the four analyses showing significant results ($p < .001$)  
- There was not a statistically significant difference between literary and informational text mean scores for either grade-level cohort  
- Females outperformed males on the informational text subtest with low to moderate effect sizes, $d = .27$ to .55  
- Student from economically advantaged backgrounds outperformed their peers from low SES backgrounds on all MAP-R subtests with low to moderate effect sizes, $d = .38$ to .6  
- Students without disabilities significantly outperformed their peers with disabilities on all MAP-R subtests with large effect sizes, $d = 1.05$ to 1.27 |

| 6. | Do MAP-R literary text and informational text scores provide unique contributions explaining variance in eMPowerME ELA/L performance for sixth-seventh and seventh-eighth graders? | Hierarchical Regression | Literary text accounted for greatest variance (53%) when entered first for the Grade 6-7; informational text accounted for an additional 3% variance ($p < .001$)  
- Informational text accounted for the greatest variance (60%) when entered first for the Grade 7-8 cohort; literary text accounted for an additional 4% variance ($p < .001$)  
- Vocabulary acquisition and use had the least amount of variance for both cohorts; failed to explain significant unique variance for the Grade 6-7 cohort and added 1% ($p < .01$) for the 7-8 cohort |
Despite the increased federal focus on reading instruction, many students in the middle-grades struggle with reading comprehension (Edmonds et al., 2009; Flagella-Luby et al., 2012; Morsy et al., 2010; NCES, 2015, 2017; Scammacca et al., 2007). The information ascertained from universal screening data can be useful for educators when identifying students who are at risk for reading problems and in need of intervention (Berkeley, Bender, Peaster, & Saunders, 2009; Brown-Chidsey, & Steege, 2010; Jenkins et al., 2007). To implement universal screening within a multi-tiered framework, schools must have psychometrically adequate measurement tools.

Schools commonly use a singular universal screener (Jenkins et al., 2013; Prewett et al., 2012) despite existing research indicating that single measures rarely result in acceptable diagnostic accuracy (Hintze & Silberglitt, 2005; Johnson et al., 2009; Kent et al., 2018). Additionally, a large body of research supports the utility of general outcome measures (GOMs) for predicting reading outcomes (e.g., Ardoin & Christ, 2008; Goffreda & DiPerna, 2010; Wayman et al., 2007). Some researchers, however, have found promising results of computer adaptive tests (CATs) as universal screening measures and, in particular, the Measures of Academic Progress (MAP) as a single predictor of reading performance on state summative assessments (January & Ardoin, 2015; Klingbeil, et al., 2017; VanMeveren, Hulac, Wollersheim-Shervey, 2018; Van Norman et al., 2017).

**A Review of Data Analyses and Results**

In this study, I examined the technical adequacy of the Measures of Academic Progress for Reading (MAP-R; Northwest Evaluation Association [NWEA], 2011; 2013; 2015a) and
demographic variables in predicting student performance on the eMPowerME English Language Arts/Literacy (eMPowerME ELA/L; Measured Progress, 2017). I also determined the minimum MAP-R scores needed in the spring of Grade 6 and 7 that demonstrated a student was on track to meet proficiency standards on the eMPowerME one year later in Grade 7 and 8. A secondary purpose was to determine whether differences existed between middle-level students’ subtests scores on the MAP-R and to examine the unique contribution of MAP-R subtests to the eMPowerME. In this remainder of this chapter, I reviewed the major findings for each research question. Additionally, I discussed the strengths and limitations of the study. Finally, I discussed notable findings and implications relative to practice and future research.

**Concurrent and Predictive Validity of MAP-R with eMPowerME ELA/L**

The first two research questions that guided this study were to determine the concurrent and predictive validity of the MAP-R and the eMPowerME ELA/L in Grades 6-7 and 7-8. Correlations were strong and statistically significant ($p < .001$) between MAP-R and eMPowerME ELA/L scores ranging from .75 to .80. That is the spring 2016 MAP-R total scores statistically explained 56% of seventh-grade and 64% of eighth-grade spring 2017 eMPowerME ELA/L scores (predictive). The spring 2017 MAP-R explained 61% of the spring 2017 eMPowerME ELA/L scores for both Grades 7 and 8 (concurrent). The Grade 6-7 cohort had a stronger correlation when the measures were taken concurrently, whereas the Grade 7-8 had a stronger correlation with scores taken one year apart. It is noteworthy that both grade-level cohorts had the strongest correlations between MAP-R and eMPowerME ELA/L during the spring of their seventh-grade year indicating there is variability in the strength of the relationships between the two tests across grade levels. Overall, there were robust correlations
between MAP-R and eMPowerME ELA/L scores for Grades 6-7 and 7-8, providing further evidence of the concurrent and predictive validity of MAP-R as a universal screening measure.

The correlational findings of the current study are consistent with previous studies. For example, January and Ardoin (2015) examined the concurrent validity between the MAP-R with the Iowa Test of Basic Skills (ITBS) for third graders and found a statistically significant correlation ($r = .87$, $p < .001$). In another study, Klingbeil et al. (2017) observed strong correlations (range = .74 to .78, $p < .01$) between MAP-R and the Smarter Balanced assessment scores in Grades 3, 4, and 5. Van Norman et al., (2017) compared MAP-R scores for fourth through seventh graders to the Minnesota Comprehensive Assessments (MCA) taken from fall to spring with correlations ranging from .70 to .82 ($p < .001$). VanMeveren et al. (2018) also investigated the relationship between MAP-R and MCA in fourth and fifth grades and found medium-to-strong associations (range = .53 to .75). The correlational evidence in the current study suggests that MAP-R captures the reading constructs tested on the eMPowerME ELA/L. This information is helpful to middle-level educators when making decisions about which universal screening measure best aligns to the state assessment, but it does not indicate a causal relationship. Technical adequacy is needed to determine the usefulness of MAP-R to predict future performance on eMPowerME ELA/L and ultimately to make high-stakes educational decisions.

The second research question was to determine the effectiveness of the MAP-R scores as an individual predictor of sixth- and seventh-graders performance (proficient or not proficient) one year later on eMPowerME ELA/L. The logistic regression models, with the MAP-R scores as the single predictor variable, were statistically significant. MAP-R explained 54.6\% and 57.8\% of the variance in proficiency and correctly classified 83\% and 80\% of cases for the
Grades 6-7 and 7-8, respectively. January and Ardoin (2015) conducted hierarchical multiple regression analyses to evaluate the unique contribution of CBM-R and MAP-R in third-grade students’ Iowa Test of Basic Skills (ITBS) performance. When MAP-R was entered into the model first, it accounted for 75% of the variance in ITBS ($p < .001$). In the current study, MAP-R had the best balance of sensitivity and specificity (sensitivity = .87 and specificity = .76) in Grade 6-7. For Grade 7-8, sensitivity (.88) was at an acceptable range, but the specificity (.67) was below a permissible level for a screener. Klingbeil et al. (2017) used logistic regression analyses using MAP-R to predict proficiency on the Smarter Balanced assessment with somewhat similar results to the current study showing variability across grade levels with Grade 4 showing a better balance of sensitivity and specificity (.86 and .83) over Grades 3 (.63 and .83) and 5 (.75 and .84). Van Norman et al., (2017) conducted logistic regression analyses with MAP-R predicting performance on the MCA in Grades 4-7 that resulted in sensitivity ranging from .79 to .88 and specificity from .81 to .86. It is important to note that sensitivity improved at increasing grade levels and the best balance of sensitivity and specificity (.88 and .84) occurred at the seventh grade. VanMeveren et al. (2018) had similar results in fourth and fifth grade with sensitivity and specificity ranging from .81 to .88 and .79 to .82.

The odds ratio in the current study produced similar results across grade levels indicating that for every point higher a student scored on the spring MAP-R as a sixth or seventh grader, the odds of achieving proficiency on the spring eMPowerME ELA/L in Grade 7 or 8 increased by a factor of 1.20 and 1.21, respectively. VanMeveren et al. (2018) reported a logistic regression model that resulted in similar odds ratios of 1.24 and 1.26 in fourth and fifth grade with MAP-R predicting performance on the MAC. The overall technical adequacy findings of this study add to the previous literature (Andren, 2010; January & Ardoin, 2015; Klingbeil et al., 2017;
Maziarz, 2010; VanMeveren, et al., 2018; Van Norman et al., 2017) suggesting that MAP-R is a strong predictor of student performance on a statewide reading achievement test in Grades 7 and 8.

**MAP-R with Demographic Variables to Predict Performance on eMPowerME ELA/L**

The third research question sought to determine if the strength of the prediction improved with demographic variables included in the model with the MAP-R scores. The results were significant with the added gender, SES, and disability factors, albeit there were some differences between grade levels when comparing the predictor variables. For instance, classification indices (e.g., sensitivity, specificity) were higher when the MAP-R scores were used as a single predictor than when combined with student demographics. The overall classification accuracy (CA) for MAP-R with added demographic variables was similar to MAP-R alone but slightly decreased for the Grade 6-7 cohort and slightly increased for the Grade 7-8 cohort. With that said, all conditional probability indices were within the acceptable range for a universal screener (Kilgus et al., 2014) for MAP-R as a single predictor variable and MAP-R with the added student characteristics.

Of the predictor variables studied, the MAP-R scores and SES were significant for the Grade 6-7 cohort and MAP-R scores and gender were significant for the Grade 7-8 cohort. Sixth graders from economically advantaged backgrounds had 2.4 times higher odds of meeting proficiency standards in seventh grade than those who were economically disadvantaged. Seventh-grade females had 3.2 times higher odds than their male peers at achieving proficiency standards on the eMPowerME in eighth grade. These findings of the current study are consistent with a previous investigation indicating that late elementary and middle school students from disadvantaged socioeconomic backgrounds are at a higher risk of developing reading problems.
compared to students from advantaged socioeconomic backgrounds (Kieffer, 2010). Additionally, gender differences showing females outperform males on reading achievement tests are congruous with the existing literature (e.g., Etmanskie et al., 2016; Klecker, 2006; 2014). Finally, the stable demographic characteristics historically associated with risk status and relevant to the sample in this study (i.e., male, low-SES, receiving special education) were overall less robust predictors of reading achievement when combined with MAP-R scores compared to the MAP-R scores alone. This finding is consistent with previous investigation indicating that baseline academic performance was a stronger predictor of performance on large-scale assessments than demographic factors (e.g., Ball, Finch, Gettinger, & Reading and Behavior Intervention Project, 2014, Ball & O’Connor, 2016).

**Optimizing MAP-R Cut Scores in Predicting Performance on eMPowerME ELA/L**

Although general CA with MAP-R alone or combined with the demographic predictor variables is essential, the ultimate goal for a universal screener is to maximize the identification of students who truly are at-risk for poor reading outcomes and in need of intervention. As such, the fourth research question was to determine the optimal cut scores on the MAP-R in predicting performance on the eMPowerME ELA/L. Area under the curve (AUC) values ranging from .80 to .90 are considered good, and values > .90 considered excellent (Catts et al., 2009; Compton et al., 2006). Both grade-level cohorts MAP-R scores resulted in high AUC values (.88 and .89) further indicating that MAP-R can be a useful screening assessment for middle schools to predict performance on eMPowerME ELA/L. Van Norman et al. (2017) and VanMaveren et al. (2018) had similar results with MAP-R predicting performance on the MCA producing AUC values ranging from .88 to .92 in Grades 4 through 7 and .89 to .92 in Grades 4 and 5.
While the results of this study added to the validity evidence of MAP-R scores in predicting performance on the eMPowerME ELA/L, the diagnostic accuracy results did not directly align with NWEA’s published norms (NWEA, 2015). For example, when a student scored below a 212.5 in Grade 7 and below 214.5 in Grade 8, the true positive rate (sensitivity) was minimized, and the true negative rate (specificity) was maximized. In other words, cut scores associated with NWEAs established risk for Grades 6 and 7 (< 40th percentile) resulted in unacceptable rates of false negatives meaning the MAP-R indicated students were average, however, later these students scored as below and well below state expectations on the eMPowerME ELA/L. These findings are consistent with existing literature indicating that single screening measures tend to produce diagnostic accuracy below the acceptable level recommended for universal screening when the publisher-provided cut-scores were used (Klingbeil et al., 2017; Van Norman et al., 2017).

By using statistically derived cut-scores, I was able to improve the diagnostic accuracy by maximizing the sensitivity and specificity to acceptable levels for a universal screener. These derived cut-scores corresponded with NWEA’s established threshold at the 60th percentile. Van Norman and colleagues (2017) produced a similar MAP-R cut score for the fall of seventh grade in predicting performance on MAC in the spring indicating that publisher recommended MAP-R cut scores for risk are set too low for at least two state tests at Grade 7. I also attained cut scores that maximized sensitivity above .90 as recommended for a universal screener by some researchers (Compton et al., 2010; Jenkins et al., 2007). These cut cores aligned with NWEA’s established norms at the 80th percentile. High sensitivity, however, resulted in lowering the specificity value below the acceptable level (.70) and increased the false positive rate indicating
MAP-R identified students as below-average who later scored as proficient on the eMPowerME ELA/L.

Overall, the above findings provide further evidence of the technical adequacy of MAP-R in predicting performance on state reading assessments but extends this to middle-level students. Specifically, the recent studies by Klingbeil et al. (2017) and VanMeveren et al., (2018) resulted in technical adequacy indices for Grades 4 and 5 that were similar to the findings of this study for Grade 6-7 and 7-8. Furthermore, the current study adds to the existing evidence indicating that locally derived cut scores improve the diagnostic accuracy from those specified by the publisher (e.g., Ball & O’Connor, 2016; Baker et al., 2015; Jenkins et al., 2007; Johnson et al., 2010, Kilgus et al., 2014; Van Norman et al., 2016).

MAP-R Performance Between and Within Subtests

The fifth question investigated whether there were any significant differences between student scores on the three MAP-R subtests and whether there were any significant differences among subtests according to student characteristics. Students in this study performed better on the MAP-R vocabulary acquisition and use (VAU) subtest compared to either the literary or the informational text subtests with three out of the four analyses showing significant results. Despite that student performance was strongest on the VAU subtest, this subtest had the weakest correlations with eMPowerME ELA/L compared to the literary text and informational text. Overall, there was not a statistically significant difference between mean scores when comparing literary and informational text for either grade-level cohort.

According to prior research, students generally struggle with comprehending informational text over literary text (Best et al., 2008; Cummins, 2013; Dennis, 2013; Eason et al., 2012; McNamara et al., 2011; Sanacore & Palumbo, 2009; Thompson et al., 2012; Zabrucky
& Ratner, 1992). The students in this study, however, performed comparably with both text types on the MAP-R assessment. This finding is of particular relevance considering the recent efforts to improve informational text comprehension by national organizations, content experts, leading researchers, and assessment developers (Duke, 2000; Jeong et al., 2010; Kamil & Lane, 1997; NAGB, 2015; NGA Center for Best Practices & CCSSO, 2010). Given the students in this sample were second and third graders when the Maine school districts began to adjust their curriculums to align with the newly adopted Common Core State Standards in English Language Arts (Maine Department of Education, 2018) it is plausible that these scores collected as sixth, seventh, and eighth graders should reflect some of these curriculum changes.

In addition to differences between MAP-R subtest scores, some notable differences occurred when analyzing scores within subtests by student demographics. For instance, females outperformed males on the literary text and informational text subtests with low to moderate Cohen’s effect size values (d = .27 to .55). Overall, this finding supports previous research reports on the NAEP and PIRLS assessments indicating the mean difference between scores for females and males were consistently higher for literary than for informational text and the achievement differences were smaller between genders for informational reading but still of practical significance (Mullis et al., 2012).

Based on SES, comparable statistically significant differences occurred between the scores of students from economically advantaged and disadvantaged backgrounds across all MAP-R subtests indicating low to moderate practical significance. This finding aligns with the recent analysis of the long-term trends of reading achievement conducted by Hanushek and colleagues (2019) that indicated the SES reading achievement gap in the U.S. has remained large and persistent over the past fifty years. Furthermore, the students with low SES from this study
performed equally as weak on literary and informational text comprehension supporting the most recent NAEP scores. That is fourth and eighth-grade students who were not eligible for the national school lunch program significantly outperformed student who qualified \( (p < .05) \) on both text types according to the 2017 results (NCES, 2019). Finally, students without disabilities significantly outperformed their peers with disabilities on all MAP-R subtests with large effect sizes ranging from 1.05 to 1.27. These findings corroborate with the meta-analysis results conducted by Gilmour and colleagues (2019), indicating that students with disabilities performed 1.7 standard deviations below their peers without disabilities on reading achievement tests.

**MAP-R Subtests as Components of eMPowerME ELA/L**

Finally, the sixth question examined the unique variance of the spring 2016 MAP-R subtests in predicting performance on the spring 2017 eMPowerME ELA/L. The combined MAP-R subtests added significant variance (56% and 64%) in explaining performance in eMPowerME ELA/L scores. The literary text and informational text subtests accounted for the highest degree of the variance in eMPowerME ELA/L scores. Although students scored significantly better on the MAP-R vocabulary acquisition and use subtest compared to literary text and informational text, this subtest had the weaker correlations with and failed to explain any significant unique variance in eMPowerME ELA/L scores one year later. These finding are consistent with prior research suggests that reading comprehension (MAP literary text and informational text) explained more variance in an outcome measure (i.e., CBM-R) for upper-elementary students than in primary students (Ardoin et al., 2013; January & Ardoin, 2015).

**Implications for School Practitioners and Policymakers**

In making sense of these findings, I took several considerations into account. The results suggested that MAP-R was strong as a single measure to predict performance on the
eMPowerME ELA/L. To the best of my knowledge, no studies exist that investigated the predictive validity between MAP-R and the eMPowerME ELA/L state assessment. Considering several middle schools in the state of Maine currently use the MAP-R test, these findings are of practical importance to administrators, curriculum leaders, and teachers. The use of MAP-R as a direct route to intervention for students in Grades 6 through 8 may be of interest to school leaders who are seeking to improve the efficiency of their screening process. From an economic standpoint, using a single screening measure may substantially reduce the expense and time spent in administering multiple assessments. In the current study, however, the MAP-R incorrectly identified 13.4% of students from the Grade 6-7 cohort and 11.9% of the Grade 7-8 cohort as at risk but who scored proficient on the eMPowerME ELA/L. Likewise, 24.1% of students from the Grade 6-7 cohort and 33.0% of the students from the Grade 7-8 were incorrectly classified as proficient on the MAP-R who later scored below proficiency on the statetest. Some administrators may view the reduced cost of administering a single screening measure to be worth the tradeoff of providing intervention to a higher proportion of students. Providing intervention to students who do not need it, might be viewed by other school leaders as unnecessarily taxing the already limited resources available for intervention. As such, these practitioners may prefer to invest the time and financial resources to administer multiple screeners in exchange for improved accuracy.

The findings of this study also underline the importance of applying locally derived cut scores (Ball & O’Connor, 2016; Jenkins et al., 2007; Johnson et al., 2010, Kilgus et al., 2014) to support educational decision making within a multi-tiered framework. Again, the MAP-R incorrectly identified 24.1% of the Grade 6-7 cohort and 33.0% of the Grade 7-8 cohort as average or above but who were not proficient on the eMPowerME ELA/L one year later. As
such, I improved the quality of classification accuracy by determining the MAP-R cut scores that maximized the number of students correctly classified as at-risk or not at-risk of meeting proficiency standards on the eMPowerME ELA/L test while minimizing the number of incorrectly classified students. These improved cut scores for the spring administration for sixth and seventh graders aligned to the scores associated with the 60th percentile according to the 2015 NWEA established norms. NWEA, however, considers scores associated with the 40th percentile and below as at risk. By using this benchmark, schools would correctly classify most all students who would score proficient on the eMPowerME ELA/L one year later but also would possibly misidentify a high number of students as being on track who will indeed not meet proficiency standards on the state test.

Overall the consideration of demographic variables combined with MAP-R scores did not substantively improve the overall classification accuracy in this study suggesting that demographics may not be useful as part of risk determination decisions. However, analyses of MAP-R subtests did shed light on notable differences in performance by student characteristics. In this study, female students outperformed male students, students from higher SES backgrounds outperformed their classmates from lower SES backgrounds, and students without disabilities outperformed their peers with disabilities confirming long-standing research on the reading achievement gaps (e.g., Reardon et al., 2018; Gilmour et al., 2019; Hanushek et al., 2019). These findings have practical implications for policymakers at the national, state, and local levels suggesting that policies over the last several decades aimed at closing the achievement gaps have had little influence on a large-scale. School teams need guidance in the implementation of tiered intervention frameworks like RtI. This study could be of practical value
to district leaders in promoting the use of assessment data as the basis for the accurate diagnosis of reading difficulties and instructional decision making.

Other notable findings regarding the difference in performance on the MAP-R subtests have practical implications for practitioners. First, students scored significantly higher on the vocabulary acquisition and use subtest compared to the literary text and informational text subtests; however, this subtest had the weakest correlations with and accounted for almost no unique variance in eMPowerME ELA/L scores. In addition to the MAP-R total score, it might be feasible for schools to use subtest scores that are specifically comprehension measures (i.e., literary text and informational text) when determining which students will not meet proficiency standards on the state test. Second, despite prior research indicating that students generally struggle with comprehending informational text over literary text (Best et al., 2008; Cummins, 2013; Dennis, 2013; Eason et al., 2012; McNamara et al., 2011; Sanacore & Palumbo, 2009; Thompson et al., 2012; Zabrucky & Ratner, 1992), the students in this study performed as equally poor with both text types. Nonetheless, given the unique differences between informational and literary text comprehension, the lack of explicit instruction on informational text structures, and the need for assessments to identify students who are struggling with comprehending informational passages, this research investigating the text types separately for unique influences on reading comprehension makes a potentially valuable contribution to the field.

Assumptions, Limitations, and Future Research

Findings from the current study add the extant literature on the validity of a universal screener in predicting performance on a high-stakes state test; however, as with all research, I must consider the assumptions and limitations of this study. In designing this study, I made three
assumptions that if violated, could threaten internal validity. First, the participating schools/districts claimed the reading curriculum aligned to the CCSS for reading, a necessary condition for the valid alignment of the school’s curriculum and the measures of reading. Second, the MAP-R and eMPowerME ELA/L test developers reported that test items corresponded to the CCSS for reading (Measured Progress, 2017; NWEA, 2011), another condition for validity. Third, I assumed that the district personnel administered the tests according to publisher guidelines; thereby ensuring the results are reliable and valid.

The sample comprised of student data from two schools in two neighboring rural districts, and participants were relatively homogenous concerning demographic characteristics. The percentage of White students attending the schools included in this study was 97%. Given the overall lack of racial and ethnic diversity in the state and the participating schools, I did not examine race as a variable. Although I attempted to reduce district- and school-specific effects by combining grade-level cohorts, additional research is needed to evaluate these findings in schools from different geographic locations and with more student diversity, including non-White students and English Learners, and urban and suburban areas. All the MAP-R data were collected during the spring universal screening window; thus, findings may not generalize to other screening time frames. Future studies should be replicated to include fall and winter administrations of MAP-R. Another limitation of this study is that the eMPowerME ELA/L scores were included for only seventh and eighth graders. Therefore, although the current investigation provided initial evidence of the concurrent and predictive validity of MAP-R with eMPowerME ELA/L, these results can be generalized only to seventh- and eighth-grade students. Given that the MAP-R is designed to be administered to students in Grades K-12 and
eMPowerME ELA/L in Grades 3 through 8, additional research is necessary to establish the validity of MAP-R with eMPowerME ELA/L at all grade levels tested.

This study broadly applies to the impact of test design on measuring reading comprehension. Although not the purpose of the present investigation, future research should continue to examine the effects of passage and question types on reading comprehension. In particular, there is a need to be able to isolate these features from one another on comprehension assessments in order to provide a more meaningful interpretation of how specific test properties affect students’ performance. Critical to the study, however, is the continuation of research on the comprehension skills measured by different standardized tests, including the eMPowerME ELA/L assessment. The purpose is not only to improve the field’s understanding of reading comprehension generally but also to learn about how specific reading assessments may differ from one another in their assessment of reading comprehension (Keenen & Meenan, 2014).

It is important to note that this study primarily focused on MAP-R as a universal screener for the purpose of identifying students at risk of not passing the state assessment, eMPowerME ELA/L. However, MAP-R data including the subtest differences within and across student characteristics may be used for many reasons such as informing instruction, evaluating intervention effectiveness, monitoring student growth, or program evaluation.

**Summary**

Universal screening is the first step in identifying students who are at risk of not meeting proficiency standards on the state test and who require additional supports. The more accurate a screener, the easier it is for education practitioners to identify those students at risk for reading difficulties and in need of intervention. The results of this study demonstrated the concurrent and predictive validity of the MAP-R in relation to seventh and eighth graders performance on the
eMPowerME ELA/L test. The demographic variables of gender, SES, and disability status did not add to the overall classification accuracy but indeed exemplified the achievements gaps among these subgroups. The students in this sample did not fare better on one type of text over the other, literary text versus informational text, despite the evidence cited in the literature indicating that informational text is generally more challenging for students than literary text comprehension. Finally, the findings of this study suggest that practitioners may benefit from applying locally derived cut scores to support their educational decision making within a multi-tiered framework. Overall, this study highlights the importance for educators to receive the proper training and supports necessary to make informed decisions around assessment practices that consider the unique characteristics of their students and make use of the available resources within their system.
REFERENCES


http://dx.doi.org/10.3102/01623737024003219


https://doi.org/10.1023/A:1008131926604


https://doi.org/10.1037/a0025323

http://dx.doi.org/10.1044/0161-1461(2001/004)

https://doi.org/10.1177/0022219408326219


Elementary and Secondary Education Act (ESEA) of 1965, Public Law No. 89-10.


Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C. § 1415 et


129


Maziarz, A. L. (2010). *The exploration of demographics and computer adaptive testing in predicting performance on state-mandated reading assessments* (Doctoral Dissertation). Retrieved from [https://pdfs.semanticscholar.org/bcdf/0862b7d3d38c17ee04ad098b598b71a2c7f0.pdf](https://pdfs.semanticscholar.org/bcdf/0862b7d3d38c17ee04ad098b598b71a2c7f0.pdf)


McNamara, D. S., Kintsch, E., Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1-43. [https://doi.org/10.1207/s1532690xci1401_1](https://doi.org/10.1207/s1532690xci1401_1)


BIOGRAPHY OF THE AUTHOR

Karen Greenwood Smith is an Instructor of Special Education at the University of Maine at Farmington since 2007. Karen teaches various undergraduate courses in Special Education that include topics on Foundations, Collaboration, Positive Behavior Interventions and Supports, Literacy Methods, and Mathematics Methods. She also has supervised practicum students in local schools. In addition to her teaching, Karen engages in many service-related activities on campus and within the local schools. For example, she is an active member of the Faculty Senate, serves as the Program Coordinator for the Special Education Department, and provides in-service teacher workshops on benchmark/progress monitoring assessments and multi-tiered intervention frameworks. Karen enjoys advising students in undergraduate research projects related to mathematics, reading, and writing interventions. Karen is looking forward to engaging with school, state, and national partners in the field and conducting research related to high-leverage practices in K-12 Special Education and pre-service teacher preparation and in-service teacher professional development. She is a candidate for the Doctorate of Education degree from the University of Maine in August 2019.