An Examination of Bias in Oral Reading Fluency:

Differential Effects across Race, Gender, and Socioeconomic Status

by

Jill Adkins

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2013 by the
Graduate Supervisory Committee:

Linda C. Caterino, Chair
Robert Atkinson
Kathryn Nakagawa

ARIZONA STATE UNIVERSITY

December 2013

ABSTRACT

Recent legislation allowing educational agencies to use Response to Intervention (RTI) in determining whether a child has a specific learning disability, coupled with a focus on large-scale testing and accountability resulted in the increasing use of curriculum based measurement (CBM) as a tool for understanding students' progress towards state standards, particularly in reading through the use of oral reading fluency measures. Extensive evidence of oral reading fluency's predictability of reading comprehension exists, but little research on differential effects across racial, gender, and socioeconomic subgroups is available. This study investigated racial, gender, and socioeconomic bias in DIBELS Oral Reading Fluency (DIBELS ORF) probes predictive and concurrent relationship with MAP reading comprehension scores for African American and Caucasian students. Participants were 834 second through fifth grade students in a school district located in a southeastern US state. The dataset consisted of student fall and spring DIBELS ORF scores and spring MAP reading comprehension scores. Concurrent correlation results between spring DIBELS ORF and MAP reading comprehension scores were moderate to large and statistically significant across all grades and demographic groups; however, correlations between fall DIBELS ORF and MAP reading comprehension scores were generally weak. Stepwise multiple regression analyses were used to examine the best variable, or combination of variables, in predicting MAP reading comprehension scores. Models differed for each grade level; however, spring DIBELS ORF scores were always included, whether alone or in combination with demographic variables, in the best prediction model. Potthoff's procedure was used to simultaneously test for slope and intercept differences among regression equations to

determine if DIBELS ORF scores from fall and spring differentially predicted MAP reading comprehension scores across demographic groups. Nine of 24 simultaneous contrasts demonstrated a significant effect; seven were related to race, one was related to gender, and one was related to socioeconomic status. Racial bias in predicting MAP reading comprehension performance from spring DIBELS ORF was found. Differential prediction among gender and SES groups was not consistent indicating little to no practical significance. Results are discussed in the context of practical implications of differential validity, both predictive and concurrent, and potential impact on disproportionality.

TABLE OF CONTENTS

Page

LIST OF TABLES v

CHAPTER

1 Introduction And Literature Review 1

Response to Intervention Defined 2

History of Curriculum-Based Measurement 8

Definition of Curriculum-Based Measurement 9

CBM-Reading 13

CBM Critiques 15

Test Bias 17

Predictive Bias of ORF Measures 19

Predictive Bias and Disproportionality 23

Purpose of the Current Study 28

Research Questions 29

2 Method 31

Participants 31

Instruments 32

Procedure 36

Analyses 37

3 Results 40

Descriptive Statistics 40

Research Question #1 41

CHAPTER Page

Research Question #2 42

Research Question #3 45

4 Discussion 47

Research Summary 47

Research Question #1 49

Research Question #2 50

Research Question #3 52

Limitations 54

Directions for Future Research 55

REFERENCES..... 57

APPENDIX

A TABLES……65

LIST OF TABLES

Table Page

A1. Demographic Characteristics by Grade 66

A2. Means and Standard Deviations for Each Measure for Grade 2 66

A3. Means and Standard Deviations for Each Measure for Grade 3 67

A4. Means and Standard Deviations for Each Measure for Grade 4 67

A5. Means and Standard Deviations for Each Measure for Grade 5 68

A6. *t* Tests Comparing Gender, Race, and Lunch Status Group Means 69

A7. Intercorrelations Among All Measures for the Grade 2 Total Sample 70

A8. Intercorraltions Among All Measures for the Grade 3 Total Sample 70

A9. Intercorrelations Among All Measures for the Grade 4 Total Sample 70

A10. Intercorrelations Among All Measures for the Grade 5 Total Sample 71

A11. Pearson Product-Moment Correlation Coefficients between DIBELS ORF and MAP Reading Comprehension Scores for Grade 2 71

A12. Pearson Product-Moment Correlation Coefficients between DIBELS ORF and MAP Reading Comprehension Scores for Grade 3 72

A13. Pearson Product-Moment Correlation Coefficients between DIBELS ORF and MAP Reading Comprehension Scores for Grade 4 72

A14. Pearson Product-Moment Correlation Coefficients between DIBELS ORF and MAP Reading Comprehension Scores for Grade 5 73

A15. Results for Stepwise Multiple Regression Analyses Predicting MAP Reading Comprehension for Grade 2 73

Table Page

A16. Coefficients for Significant Predictor Variables for Multiple Regression Analysis Predicting MAP Reading Comprehension for Grade 2 73

A17. Results for Stepwise Multiple Regression Analyses Predicting MAP Reading Comprehension for Grade 3 74

A18. Coefficients for Significant Predictor Variables for Multiple Regression Analysis Predicting MAP Reading Comprehension for Grade 3 74

A19. Results for Stepwise Multiple Regression Analyses Predicting MAP Reading Comprehension for Grade 4 74

A20. Coefficients for Significant Predictor Variables for Multiple Regression Analysis Predicting MAP Reading Comprehension for Grade 4 74

A21. Results for Stepwise Multiple Regression Analyses Predicting MAP Reading Comprehension for Grade 5 75

A22. Coefficients for Significant Predictor Variables for Multiple Regression Analysis Predicting MAP Reading Comprehension for Grade 5 75

A23. *F*, *df*, and *p* Values for Simultaneous Slope and Intercept Comparisons Between Demographic Groups in Predicting MAP Reading Comprehension Scores Using Fall DIBELS ORF and Spring DIBELS ORF for Grade 2 76

A24. *F*, *df*, and *p* Values for Simultaneous Slope and Intercept Comparisons Between Demographic Groups in Predicting MAP Reading Comprehension Scores Using Fall DIBELS ORF and Spring DIBELS ORF for Grade 3 77

Table Page

A25. *F*, *df*, and *p* Values for Simultaneous Slope and Intercept Comparisons Between Demographic Groups in Predicting MAP Reading Comprehension Scores Using Fall DIBELS ORF and Spring DIBELS ORF for Grade 4 78

A26. *F*, *df*, and *p* Values for Simultaneous Slope and Intercept Comparisons Between Demographic Groups in Predicting MAP Reading Comprehension Scores Using Fall DIBELS ORF and Spring DIBELS ORF for Grade 5 79

Chapter 1

Traditionally, the identification of students with specific learning disabilities has largely relied on the documentation of a significant discrepancy between cognitive ability and academic achievement (IQ-achievement discrepancy) in one or more of the following academic skill areas; oral expression, listening comprehension, written expression, basic reading skills, reading comprehension, mathematics calculation, and mathematics reasoning, with reading fluency added in 2004 (Busch & Reschly, 2007; Vaughn & Fuchs, 2003). Despite widespread use of the IQ-achievement discrepancy, a considerable amount of criticism surrounds this method of SLD identification. As a result, practitioners and legislators have sought alternative approaches to SLD identification.

In addition to the IQ-achievement discrepancy method of SLD identification, the Individuals with Disabilities Education Improvement Act (2004) permits the use of alternative, research-based approaches in determining SLD. This alternative method of SLD identification, which is largely grounded in the Cattell-Horn-Carroll (CHC) theory of cognitive abilities, requires the identification of specific and statistically significant academic and cognitive strengths and weakness, as well as average or above average intelligence (Flanagan, Fiorello, & Ortiz, 2010). It is presumed that an observable and meaningful relationship exists between cognitive deficits and academic deficits in students with SLD such that the cognitive deficit is the presumed cause of the academic deficit (Flanagan et al., 2010). Several different models of this approach exist, each of which share three common components (Flanagan et al., 2010). First, cognitive strength is demonstrated by average or higher abilities and processes. The second common component is the presence of academic weakness or failure that is

unexpected because of overall cognitive ability that is at least average. This difference between overall cognitive ability and academic skill must be statistically significant. The final common component is a documented cognitive deficit that is specific, because overall cognitive ability is at least average. This difference between overall cognitive ability and specific cognitive deficit, or processing deficit, must also be statistically significant.

Response to Intervention (RTI) is a third method of SLD identification. Section 1414(b)(6)(B) of the Individuals with Disabilities Education Improvement Act, signed into law by President Bush in 2004, indicates that "In determining whether a child has a specific learning disability, a local educational agency may use a process that determines if the child responds to scientific, research-based intervention as part of the evaluation process" (IDEA 20 U.S.C. § 1414(b)(6)(B)). As a result, many school districts are now using Response to Intervention (RTI) as a substitute for, or supplement to, the other two standardized assessment models to identify students with. RTI is also a means of identifying students in need of, and providing early intervention to, all children at risk for school failure.

Response to Intervention Defined

One of the underlying premises of RTI is the possibility that a child's struggles may be due to inadequate curriculum or instruction either in use at the present time or in the child's past (National Dissemination Center for Children with Disabilities [NICHCY], 2012). Theoretically, applying scientific, research-based intervention to academic deficits allows practitioners to rule-out inadequate curriculum or instruction as the main factor affecting performance. In simplest terms, RTI is a process by which

students are provided with quality instruction, their progress is monitored, those who do not respond receive additional instruction, and the cycle begins again until the student is performing at grade level or the child is considered for special education. Depending on state and district guidelines, students who still do not respond either qualify for special education or are referred for a special education evaluation (Fuchs, Mock, Morgan, & Young, 2003). Students' severe educational need, coupled with a lack of educational benefit, or lack of response, from high-quality interventions may be considered a sufficient condition for determining eligibility in an RTI approach to SLD identification (Shinn, 2007).

The basic concept of RTI is that students can be provided with effective interventions and information about their response, or lack thereof, can be used to guide service delivery decisions (VanDerHeyden, Witt, & Gilbertson, 2007). Response to Intervention can be further defined by a set of guiding principles. These include a multi-level prevention system, universal screening, progress monitoring, and data-based decision making (Fuchs & Fuchs, 2005, 2006; Shinn, 2007).

The Response to Intervention system includes multiple tiers, or levels of intensity or prevention, typically ranging from two to four tiers (Fuchs et al., 2003). At each tier, the intensity of academic intervention increases through practices such as more systematic and explicit instruction, increased frequency or duration of intervention, smaller groups of students, or assigning teachers with greater expertise to higher tiers of intervention (Fuchs & Fuchs, 2006). The primary prevention level is high quality, core academic instruction provided to all regular education students. The secondary level includes the addition of evidence-based intervention of moderate intensity. Higher levels

include individualized intervention of increased intensity for students who show minimal response, or lack of response, to secondary level intervention (National Center on Response to Intervention, 2010).

Universal screening, a second core feature of response to intervention systems, is conducted to identify students who may be at risk for poor learning outcomes. Universal screening tests are conducted with all students using brief academic skill measures. Once tiered intervention is in place, progress monitoring is used to quantify rate of improvement, or responsiveness to instruction, and to evaluate the effectiveness of instruction. In progress monitoring, as in universal screening, the importance of fidelity of implementation and selection of evidence based tools, with consideration for cultural and linguistic responsiveness and recognition of student strengths is emphasized (National Center on Response to Intervention, 2010). Both universal screening and progress monitoring require tools that are technically sound and enable educators to make informed decisions about student progress over time (Busch & Reschly, 2007). Shinn (2007) further argues that the quality of progress monitoring tools be no less than the quality of intervention; meaning that progress monitoring tools must also be scientifically based.

Data-based decision making is a final component of any RTI system and occurs at all levels of implementation and instruction. School teams use screening and progress monitoring data to make decisions about instruction, movement within the multi-level prevention system, and disability identification in accordance with state laws (National Center on Response to Intervention, 2010).

Despite this core set of guiding principles, implementation of RTI differs among the many states and school districts that implement the process. Some of the most noteworthy differences are the number of levels of the process, personnel who deliver interventions, and whether the process is a precursor to a formal evaluation for special education eligibility or if RTI itself is the eligibility evaluation (Fuchs et al., 2003). Despite the range of differences in implementation, two main models of RTI have emerged from the literature: the problem-solving model and the standard protocol model.

The problem-solving model is a more flexible process with emphasis on individualized interventions. Problem-solving teams conduct systematic analysis of instructional and environmental variables, determine target skill/subskill deficits, and design individualized and targeted interventions (Christ, Burns, & Ysseldyke, 2005). This model assumes that effective intervention cannot be determined prior to the systematic analysis of individual student variables. It further assumes that no single intervention will be effective for all students of a particular group. Instead, solutions to academic skill deficits are induced by evaluating students' responsiveness to a four-stage process: problem identification, problem analysis, plan implementation, and problem evaluation (Fuchs et al., 2003).

During problem identification, the first stage of the problem-solving approach, the major objective is to define the problem in concrete and observable terms. Additionally, a baseline measure of performance is obtained. In the problem analysis stage a plan is developed to address the instructional and student variables identified in the problem identification stage. Next, the plan is implemented as designed by the problem-solving team. Finally, the effectiveness of the intervention is continually evaluated and modified,

if needed. Successful solutions are often achieved after intervention modification; consequently, the problem-solving model has been dubbed the trial-and-error approach.

At each problem solving level, the process is meant to be the same: problem-solving teams determine the magnitude of the problem, analyze possible causes, design and conduct goal-directed interventions, monitor student progress, modify interventions as needed based on student responsiveness, evaluate intervention effectiveness and plan for future actions (Fuchs & Fuchs, 2006). As the intensity of student needs increase at each level, so do the educational resources and expertise employed by the problem-solving team (Fuchs et al., 2003).

When the standard protocol model is implemented, a standard set of empirically supported instructional approaches, or interventions, are implemented with the intent of preventing and remediating academic problems (Christ et al., 2005). Where the problem-solving approach is individualized for each child, the standard protocol model is not (Fuchs & Fuchs, 2006). In the standard protocol approach, the same empirically validated treatment is provided to all students experiencing problems in a given academic domain (Fuchs et al., 2003).

The fundamental difference between the standard protocol approach and problem-solving model is the level of individualization and the depth of problem analysis that occurs prior to the selection, design, and implementation of intervention (Christ et al., 2005). Through the problem-solving model, an effort is made to personalize assessment and intervention making this model more sensitive to individual student differences (Fuchs et al., 2003). Fuchs and Fuchs (2006) argue that this individualization also represents a potential weakness of the problem-solving model because it presumes

extensive expertise on the part of practitioners and problem-solving team members. Despite this presumption of considerable expertise, the problem-solving model is favored over the standard protocol model by most practitioners. In contrast, researchers favor the standard protocol model (Fuchs & Fuchs, 2006). Some distinct advantages of the standard protocol model over the problem-solving model have been noted. The standard protocol model enables greater quality control because it is easier to train practitioners to conduct one intervention correctly and to assess accuracy of implementation of one intervention. The efficacy of intervention may be assessed more easily since no other variables are involved. Additionally, when individualization of intervention program is removed, a larger number of students are able to participate in a generally effective treatment protocol (Fuchs et al., 2003).

Policymakers are hopeful that RTI will provide practitioners with solutions to the problems presented by the IQ-achievement discrepancy model. RTI has been documented to provide more assistance more quickly to a greater number of children at risk for school failure. RTI represents a valid method of SLD identification because providing individualized, intensive instruction to low performing students effectively separates students with disabilities from those who perform poorly because of inadequate prior instruction. This distinction between truly learning disabled children and children who perform poorly due to inadequate instruction leads to a reduction in special education enrollment and, consequently, cost (Fuchs & Fuchs, 2005).

Despite the inclusion of all three methods of SLD in IDEIA (2004) and in the accompanying federal regulations (34 CFR 300.540-543), each has been scrutinized in the literature and no one method in isolation has been deemed best practice for the

identification of SLD. In fact, some proponents have emerged advocating for a hybrid of both RTI and comprehensive assessment models for SLD identification where students presenting with learning difficulties are served through a RTI system, but comprehensive evaluation of the basic psychological processes following failure to respond occurs (Fuchs et al., 2003; Hale, Kaufman, Naglieri, & Kavale, 2006). Hale et al. (2006) maintain that this "balanced practice model" addresses both the definitional criteria and the method for determining SLD eligibility posed by IDEIA (2004). Suffice it to say, the use of RTI practices, whether in isolation or in combination with comprehensive evaluation approaches, now plays a major role in the identification of SLD in the United States.

In order to deliver appropriate and effective intervention to students in need, as required in any RTI or hybrid model, a consistent and accurate screening system for identifying those students is essential (Hosp & Ardoin, 2008). In addition to the need for accurate identification, it is also essential to accurately and consistently measure a student's response to the provided intervention. Essential to an RTI or hybrid model of SLD identification is the availability of measures that are technically adequate, can be administered frequently, and are sensitive to student growth (Busch & Reschly, 2007). Curriculum-based measurements (CBM), sets of procedures for measuring academic proficiency in the basic skill areas of reading, math, spelling, and written expression (Deno, 1985), serve as the measure for identification and progress monitoring.

History of Curriculum-Based Measurement

Deno and Mirkin originated the idea of CBM in 1977 at the University of Minnesota Institute for Research on Learning Disabilities in order to test the effectiveness

of a special education intervention program, called data-based program modification (DBPM; Deno & Mirkin, 1977). Deno & Mirkin's (1977) DBPM model was based on the hypothesis that formative evaluation used in a repeated manner could be used to evaluate and drive instructional methods for special education students. This research on DBPM led to the establishment of progress monitoring procedures for reading, spelling, and written expression that met acceptable standards for technical adequacy, treatment validity or utility of the measures, and logistical feasibility (Deno, 2003a). The results of this research, and the progress monitoring procedures developed as a result, laid the foundation for the assessment approach known as curriculum-based measurement (Deno, 2003b).

Definition of Curriculum-Based Measurement

When material for assessment is drawn directly from the instructional materials used by teachers in the classroom, the approach is broadly referred to as curriculum-based (Deno, 2003b). Curriculum-based assessment (CBA), as opposed to curriculum-based measurement, is the term used to refer to this wide range of informal assessment procedures. In the broad sense, curriculum-based assessment is the common process of gathering information about students' performance in the curriculum for the purpose of decision making and includes practices such as grading worksheets, calculating percentage correct, conducting error analyses of oral reading from text, or determining mastery via an end of unit test while curriculum-based measurement is a distinct subset of CBA that separates measurement materials from the curriculum, while retaining instructional relevance and allowing for technical adequacy (Deno, 2003a; Fuchs & Deno, 1994).

Hintze, Christ, and Methe (2006) describe CBA as the "umbrella" term under which many different CBA practices fall. At the next level down, CBA practices can be divided into two groups based on test-specification practices (Fuchs & Deno, 1991). Representing the majority of assessments under the CBA umbrella is specific subskill mastery measurement, where criterion-referenced assessment items are designed to gauge mastery of individual subskills, or objectives, within the broad curriculum. Specific subskills mastery measurement allows for the assessment of whether or not a certain level of mastery has been attained with one particular aspect of the curriculum, rather than assessment of skill development across an entire curriculum (Hintze et al, 2006).

The second subset of CBA, as defined by Fuchs and Deno (1991), is general outcome measurement (GOM). Curriculum-based measures are examples of general outcome measures (Silberglitt & Hintze, 2005). As the development and refinement of CBM has occurred over the last few decades, it has been substantiated that assessment materials drawn from sources other than the direct instructional materials used in the classroom by teachers provide technically adequate and instructionally relevant data (Fuchs & Deno, 1994). When material for assessment is drawn from alternative sources, rather than directly from the instructional materials used by teachers in the classroom, the assessments are referred to as general outcome measures (GOM's) or dynamic indicators of basic skills (DIBS) (Fuchs & Deno, 1994; Shinn, 1995). This separation of CBM from a school's curriculum made it possible to standardize stimulus materials while retaining the relevance of CBM for instructional decision making (Deno, 2003a).

Curriculum-based measurement is a distinct subset of CBA that refers to the specific set of formative evaluation procedures for measuring student growth in basic

skills that resulted from the research by Deno and colleagues in the 1970's (Deno, 2003a). Its focus on broad goals of a curriculum, rather than mastery of short-term objectives, allows for assessment of the retention and generalization of learning across time (Hintze et al., 2006). CBM is described as dynamic, as it is sensitive to the short-term effects of instruction and has the ability to assess change over time since the same performance objective is continually assessed (Hintze et al., 2006). In sum, CBM is considered to be simple, reliable, valid, and can be used frequently and repeatedly to measure growth.

Curriculum-based measurement is further defined and differentiated from the broader CBA by several essential characteristics. Specified measurement and evaluation procedures are delineated for CBM, including methods for generating test stimuli, administration and scoring procedures, and methods for summarizing and making inferences from data collected (Hintze et al., 2006). In addition to these defining characteristics, CBM also offers unique characteristics such as cost-effectiveness and efficiency.

Reliability and validity of CBM have been achieved through the use of standardized observational procedures for repeatedly sampling performance on core reading, writing, and mathematical skills (Deno, 2003b). The measurement tasks of CBM (e.g., spelling, oral reading fluency) are empirically selected and, therefore, reflect whether the instruction directly results in improvement in general reading outcomes (Deno, 2003a). This process of developing CBM procedures increases the criterion validity of CBM measures. Tasks selected for use in CBM are those for which reliable

measures can be constructed. This establishment of reliability includes inter-observer agreement, test/retest, reliability, and alternate form reliability (Deno, 2003a).

Standard administration and scoring procedures are specified for CBM that detail duration of the measurement, student directions, and scoring procedures (Deno, 2003b). Such standardization of the measures allows for increased reliability as well as expanding the use of data for individual and group comparisons over time (Deno, 2003b).

Formative evaluation used in a repeated manner is the crux of CBM. Obtaining repeated samples of student performance on equivalent forms of the same task across time is required to measure change. When an increase or decrease in CBM performance is measured via repeated CBM administration, that change is interpreted as a generalizable change in skill proficiency (Deno, 2003a). Each repeated measurement of CBM must be in response to a stimulus task that is unfamiliar to the student so that any increase in performance represents real growth in general proficiency rather than practice effects (Deno, 2003a). Thus, multiple forms of the stimulus task must be available which are equivalent in the basic skill measured, as well as the difficulty level of that skill. Task difficulty is held constant so that inferences regarding generalizability of student proficiency may be drawn (Deno, 2003b).

Additional characteristics of CBM relate to the efficiency with which the measures are used and the economical practicality. Frequent, repeated samples of student performance are required to measure growth. To accommodate this necessity, CBM tasks are short in duration and, therefore, do not disrupt instructional time. Moreover, because CBM material production is inexpensive, many forms can be made available for frequent,

repeated sampling. Finally, CBMs are easy to teach allowing the procedures to be used in such a way that the data are reliable (Deno, 2003a).

CBM-Reading

The initial purpose of CBM was to aid special education teachers in evaluating the effectiveness of their reading, spelling and written expression instruction (Deno, 2003a). Since the idea of repeated formative evaluation was originated by Deno and Merkin in 1977, expansions in the application of CBM have become far reaching. The expansions include use with both general and special education populations (Keller-Margulis, Shapiro, & Hintze, 2008), including deaf populations (Deno, 2003a) and English Language Learners (Deno, 2003a; Reschly, Busch, Betts, Deno, & Long 2009). The use of CBM has been extended to other age groups, including infants, preschoolers, kindergartners, and middle and secondary students (Reschly et al., 2009). CBMs have been translated into other languages and used in other countries (Reschly, et al., 2009). Content areas assessed have expanded to include social skills, pre-academic skills, mathematics, and vocabulary (Reschly et al., 2009). Additionally, the utility of CBM has expanded beyond measuring student progress to include screening and eligibility for interventions and special education services (Keller-Margulis et al., 2008), instructional placement and progress monitoring (Keller-Margulis, et al., 2008), evaluating the reintegration of special education students into regular education classrooms (Reschly et al., 2009), creation of school and district norms (Reschly et al., 2009), program evaluation (Reschly et al., 2009), universal screening (Hosp, Hosp, & Dole, 2011), and predicting success on high-stakes assessment (Deno, 2003a).

Undoubtedly, the most widely researched and utilized CBM is the oral reading measure, hereafter referred to as R-CBM (Reschly et al., 2009). For the R-CBM, students are given a passage at their grade or instructional level and are asked to read aloud from the passage for one minute. The passages are then scored for number of words read correctly, which provides an index of the student's reading fluency (Reschly et al., 2009).

Researchers define reading fluency as the rate and accuracy of oral reading in connected text (Fuchs & Fuchs, 1992; Hasbrook & Tindal, 2006; Shinn, Good, Knutson, Tilly, & Collins, 1992). The key reason for focusing on the development of reading fluency is the relationship between reading fluency and comprehension, the end goal of reading (Meyer & Felton, 1999). Fluent, or quick and accurate reading, allows the reader to attend to the meaning of text rather than to the mechanics of reading (Adams 1990; Samuels 1979). This relationship is supported by empirical research demonstrating strong correlations between reading fluency and comprehension (Shinn et al., 1992). Presumably, growth in reading fluency, as measured by R-CBM across time, indicates that, overall, a student is becoming a better reader (Reschly et al., 2009).

R-CBM scores have been evaluated according to traditional psychometric criteria for reliability and validity and have been found to demonstrate technical adequacy as a measure of reading fluency (Marston, 1989). Correlations between measures of oral reading fluency and both published measures of reading fluency and state reading assessment are consistently moderate to strong (Baker et. al, 2008).

Implementation of the No Child Left Behind Act in 2001 placed a focus on large-scale testing and accountability (NCLB, 2001). As a result, CBM has become increasingly more significant as a standardized measurement tool for understanding

students' progress towards and achievement of state standards, particularly in reading. The unique features of R-CBM, including its psychometric properties, ability to function as a general outcome measure, and the ease of administration, time efficiency, low cost, and frequency with which the measures may be given, has led to widespread use in U.S. schools (Reschly, et al., 2009). These same properties make R-CBM worthy of analysis as a direct measure of reading fluency and as a correlate to reading comprehension and general reading proficiency.

Multiple CBM systems are available to assist schools in monitoring students' acquisition of reading skills and most include one minute oral reading fluency measures. One such example is the System to Enhance Educational Performance (STEEP; Witt, 2007). Perhaps the most widely adopted of all R-CBM measures is Dynamic Indicators of Early Literacy Skills Oral Reading Fluency (DIBELS ORF). The Reading First guidelines of No Child Left Behind (2002) require states seeking federal Reading First grant funds to incorporate assessment programs that directly evaluate phonemic awareness, phonics, fluency, vocabulary, and comprehension. The DIBELS assessment system is one approved by Reading First to assess these skill areas. Because the 2002 Reading First guidelines mandate fluency instruction and assessment, reading fluency has risen to a high level of prominence, as has the use of DIBELS ORF to assess fluency and group students for intervention and instruction.

CBM Critiques

Despite the previously mentioned positive characteristics of curriculum based measures, some researchers have questioned the utility of CBM, and of DIBELS in particular. One major criticism concerns the nature of words correct per minute measures,

such as R-CBM and DIBELS ORF, and their relation to comprehension. It has been argued that the relationship between reading fluency and reading comprehension is developmental in nature, meaning the relationship changes as children age (Valencia, Smith, Reece, Li, Wixson, & Newman, 2010). Per Valencia et al. (2010), when children are acquiring decoding skills and automaticity at younger ages the relationship between reading fluency and comprehension is stronger than at older stages when these decoding and automaticity skills are more fluent and more focus is on comprehension. Therefore, Valencia et al. (2010) argue that reading fluency may not be a good indicator of reading comprehension across all ages.

The timed nature of curriculum based measures has also been called into question for the reason that timed tasks may disadvantage some readers and advantage others. Goodman (2006) notes that readers who are cautious, thoughtful, curious, talkative, or just slow are more likely to suffer in a timed test. Those who are eager, frenetic, impetuous, or drilled for the tests are likely to be advantaged in a timed test. Further, some available information suggests that measures of rate taken over very short durations may result in an overestimation of rate. The National Assessment of Educational Progress 2002 Special Study of Oral Reading indicated that students read at a faster rate for the first minute of oral reading than across the remainder of an entire 198-word passage, yet most CBM's are administered for just one minute (Daane, Campbell, Grigg, Goodman, & Oranje, 2005). The majority of reading that students perform requires considerably more sustained effort and time than does a one-minute reading sample (Valencia et al., 2010). Further, curriculum based fluency measures may not be a particularly good indicator of a student's ability to analyze more sophisticated literature or to learn new information from

complex expository texts encountered in the later grades (Valencia, et al., 2010). In general, these arguments question whether such assessments can reliably predict children's ability to read and comprehend non-test reading material and authentic texts (Goodman, 2006; Shelton, Altwerger, & Jordan, 2009).

The use of cut scores or benchmarks for determining a reader's risk, such as those recommended by DIBELS, has also been called into question. Valencia et al. (2010) indicate that the use of benchmarks misidentifies a substantial percentage of students. Misidentification results in both false negatives (the failure to identify students at risk who are at risk) and false positives (the identification of students as at risk who are not at risk). Consequently, intervention may not be provided to students in need or limited resources are wasted on students who do not require them.

Finally, it has been suggested that DIBELS and other R-CBM's are based on a flawed theory of reading because these assessments attend to discrete, or constrained, skills (Goodman, 2006; Shelton, Altwerger, & Jordan, 2009). Per Goodman, in this reductionist theory of reading, too great a focus is placed on the parts of reading rather than the "orchestrated whole of reading as a skilled human process" (2006, p. *xi*). He further argues that when the component skills are reduced to a task that can be tested in a minute only a reduced aspect of the skill is actually tested. In the case of oral reading fluency, for example, he notes that only speed and accuracy are tested and that the ability to make sense of connected text is ignored (Goodman, 2006).

Test Bias

The term bias takes on numerous different connotations that vary greatly among the general public and researchers. The term is often confused with, or used instead of,

offensiveness or fairness. Even within the scientific literature, bias goes by many names and has many characteristics; however, bias always involves scores that are too low or too high to accurately represent or predict an individual's skills, abilities, or traits (Reynolds & Ramsay, 2003). Jensen (1980) argued that test bias, separate from test fairness, is an empirically based statistical issue concerning the psychometric properties of a test as used with two or more subpopulations. Statistical techniques are necessary to detect this test bias. "In statistics, bias refers to systematic error in the estimation of a value. A biased test is one that systematically overestimates or underestimates the value of the variable it is intended to assess. If this bias occurs as a function of a nominal cultural variable, such as ethnicity or gender, cultural test bias is said to be present" (Reynolds & Ramsay, 2003, p. 68).

Tests may be biased in their content validity, construct validity, or predictive validity. Tests are biased in content validity if items behave differently for individuals of different groups. Items may be said to contain content bias if the solution required is unfamiliar to a particular group of examinees or if a particular group of examinees are penalized for providing responses that are correct in their own culture, but not in the culture for which the test was designed (Reynolds & Ramsay, 2003). Tests are biased in construct validity if they measure different traits, or constructs for individuals of different groups, or if they measure the same trait with a different degree of accuracy (Reynolds, 1982). Of these three types mentioned, issues of predictive validity are most important when dealing with the practical use of test scores in making educational selection decisions (Brown, Reynolds, & Whitaker, 1999).

Predictive validity is defined as the effectiveness of a test in predicting an individual's performance in specified activities (Anastasi, 1988). Jensen (1980) defined predictive bias as "systematic error (as contrasted to random errors of measurement) in the prediction of a criterion variable for persons of different subpopulations as a result of basing prediction on a common regression equation for all persons regardless of their subpopulation membership…" (p. 380). When one regression equation is incorrectly used for two or more groups, predictive bias occurs (Reynolds & Ramsay, 2003).

Others have defined predictive bias in similar terms: Cleary, Humphreys, Kendrick, and Wesman (1975) defined predictive bias as constant error in prediction, or error in prediction that exceeds the smallest feasible random error, as a function of group membership. The regression equation must be the same for all groups. Significant differences in slope or intercept would indicate that a single regression equation for all groups would predict inaccurately and that bias has been found (Reynolds, Lowe, & Saenz, 1999).

Predictive Bias of ORF Measures

In 1974, LaBerge and Samuels theorized that reading automaticity, or oral reading fluency, is directly related to reading comprehension. Since that time, extensive research in both general and special education has documented support for the use of oral reading fluency as a measure of reading comprehension (Baker et al., 2008).

One of the first studies to examine racial/ethnic and gender bias on oral reading fluency found that oral reading fluency passages are biased predictors of reading comprehension (Kranzler, Miller, and Jordan, 1999). A randomly selected sample of 326 Caucasian and African American students in grades 2 through 5 was administered grade

level oral reading fluency passages and the Reading Comprehension portion of the California Achievement Test (CAT). A series of multiple regression analyses were conducted by grade level. In grades 4 and 5, intercept bias was found. In grade 5, both slope and intercept bias was found for Caucasian and African American students. Oral reading fluency measures overestimated the reading comprehension of African American students and underestimated the reading comprehension of Caucasian students.

With the intention to replicate and extend the work of Kranzler, et.al (1999), Hintze, Callahan, Matthews, Williams, and Tobin (2002) examined the differential predictive bias of oral reading fluency across 136 African American and Caucasian second through fifth grade students. Their results were in direct contrast to the Kranzler et al. (1999) results. The outcome of a series of multiple regression analyses indicated that African American and Caucasian students did not differ significantly with respect to slope or intercept compared to the overall group prediction. Also, when compared directly, neither group differed significantly in slope or intercept. Oral reading fluency neither over- or under-predicted reading comprehension skills controlling for age, sex and socioeconomic status.

One major differentiation between the Kranzler et al. (1999) study and the Hintz et al. (2002) study is that Hintze et al. (2002) accounted for the developmental effects of reading by including age in the regression model. To do this, Hintze et al. (2002) used the same third grade CBM reading passage for all second through fifth graders who participated in the study. Hintze et al. (2002) noted that without age as a developmental indicator entered into analyses, all other variables have an increased chance of accounting for significant portions of variability in the criterion measure due to chance. Kranzler et

al. (1999) did not account for developmental effects in this manner which may assist in explaining the difference in results of the two studies.

Using both simultaneous multiple regression and stepwise regression procedures, Hixson and McGlinchey (2004) assessed economic and racial bias for 442 students in fourth grade using oral reading fluency scores to predict comprehension on two group measures of reading comprehension; Metropolitan Achievement Tests, Seventh Edition (MAT/7) and Michigan Educational Assessment Program (MEAP), a state reading test. The simultaneous multiple regression resulted in a significant contribution of racial group, free lunch status, and CBM ORF on MAEP scores; each of the three variables also contributed significantly to MAT/7 scores in the simultaneous multiple regression indicating that CBM ORF scores used alone are biased predictors of MAEP and MAT/7 performance. The nature of this bias was examined further and indicated evidence of intercept bias for SES and race for the MAEP: no slope or intercept bias was found for the MAT/7. Despite the significant difference in intercepts between racial groups and lunch status groups on the MAEP, the difference in predictions based on the common regression line from those based on the group membership lines was small. No bias in predicting MAEP or MAT/7 performance was found using the stepwise regression procedure. The authors stated that the nature of MAEP bias cannot be concluded from their study, but they did offer two possibilities. First, free lunch status, a dichotomous variable, was used as the indicator of SES. Use of a continuous predictor variable may have accounted for a greater portion of variance in MAEP scores and, therefore, the contribution of race may have been reduced to a non-significant amount. Second, they suggest that the MAEP test itself may be biased. Taking all analyses into account, the

authors concluded that the contribution of SES and race added very little to the prediction of reading comprehension scores. Further, they stated that although evidence of bias in CBM ORF predicting reading comprehension performance was found, the practical implications of such may be trivial.

A few additional studies have documented effects of predictive bias among ethnicities other than African Americans and Caucasians and language backgrounds other than English. Bias for ethnicity, gender, language background, and socioeconomic status was examined among a sample of nearly 4,000 Caucasian and Hispanic students in grades one through three (Klein & Jimerson, 2005). A series of hierarchical multiple regression analyses was conducted with oral reading fluency predicting Stanford Achievement Test-Ninth Edition (SAT-9) scores. Intercept bias was found; however, results indicated that the combination of factors, and not any one factor in isolation, contributed significantly to intercept bias.

Intercept bias was also documented among a sample of 543 Caucasian and Native American students when oral reading fluency scores were used to predict reading comprehension performance on the Dakota State Test of Educational Proficiency (DStep), a state measure of adequate yearly progress (Pearce & Gayle, 2009). Although oral reading fluency was found to be a robust predictor of reading comprehension for both Caucasian and Native American cohorts, significant differences were found between the separate predictive models indicating it may be best to use separate models for Caucasians and Native Americans in predicting reading comprehension performance.

Hosp et al. (2011) examined DIBELS Oral Reading Fluency and Nonsense Word Fluency for evidence of bias in predictive validity among the disaggregation categories of

the No Child Left Behind Act (economic disadvantage, limited English proficiency, disability status, and race/ethnicity) using a sample of 3,805 first through third graders through use of Receiver Operating Characteristic (ROC) curves and quantile regression. Results, similar to studies using multiple regression analyses, indicated that bias in predictive validity was found to vary by grade and disaggregation category. Of note, African American students were removed from the study due to low numbers in the sample.

When the existing body of research is examined as a whole, it is apparent that no clear pattern of differential prediction has been consistent across ethnicity, gender, or grade level (Hosp et al., 2011). Currently, caution in use of oral reading fluency with diverse students is warranted. The continuation of rigorous examination of possible bias with ethnic, gender, and socioeconomic status groups through diverse psychometric techniques is recommended (Reynolds & Ramsay, 2003).

Predictive Bias and Disproportionality

A majority of research on predictive bias has focused on major ability and aptitude tests. This research has largely shown a lack of evidence of predictive bias (Reynolds & Ramsay, 2003). In contrast, curriculum-based measures used for universal screening are often characterized by high rates of under- or over-identification which have been shown to differentially affect different subgroups of students (Cleary et al., 1975; Hosp et al., 2011). Brief screening measures, such as R-CBM and other CBMs, tend to have low reliability compared with major ability and aptitude tests: low reliability, in turn, may lead to bias in prediction (Reynolds et al., 1999). Because CBM's are widely used for both identification for remediation programs in regular education and the

identification of students with Specific Learning Disabilities, possible predictive bias in CBMs may contribute significantly to disproportionality in special education. Predictive bias in screening instrumentation may contribute to inequitable provision of remediation programs provided through general education, impacting educational achievement, and consequently, increasing the risk for special education referral, ultimately contributing to the disproportionate representation of minority student in special education programs (Skiba et al., 2008).

High rates of over- or under-identification via screening measures are often implicated in the disproportionate representation of minority students in special education (Hosp & Reschly, 2003). Since much of the value of a screening measure is determined by its ability to predict future outcomes on a criterion measure, the extent to which the inferences of future performance hold true for all subpopulations of interest is an essential area of investigation (Betts et al., 2008). Given the increased emphasis on assessment and accountability, the influence of assessment on student outcomes and the importance of examining bias in predictive validity have never been higher (Hosp at al., 2011).

However, the research on the predictive validity of criterion referenced measures is limited. While there is extensive evidence of oral reading fluency's predictability on measures of reading comprehension, there is little research on the differential prediction, or predictive bias, of racial or ethnic subgroups. The predictive bias research for major ability and aptitude tests indicates that when group differences in regression formulas are present, criterion scores of minority groups are generally over-predicted (Brown et al., 1999; Reynolds & Ramsay, 2003). Brief screening measures, such as R-CBM and other

CBMs, tend to have low reliability compared with major ability and aptitude tests: low reliability, in turn, may lead to bias in prediction (Reynolds et al., 1999). Reynolds and Ramsay (2003) note that these over-predictions of major ability and aptitude tests do not likely account for undesirable placements or diagnosis of these groups. However, an over-prediction of minority groups on a criterion measure may result in educational agencies failing to provide regular education interventions and may potentially lead to the under-identification for compensatory programs for minority students due to the fact that screening instruments over-predict their actual reading comprehension skills. Conversely, for non-minority students, screening measures could potentially over-identify the need for compensatory remediation in reading because their performance on screening measures may underestimate their true reading comprehension abilities (Hintze et al., 2002).

Disproportionality has been defined as "the representation of a group in a category that exceeds our expectations for that group, or differs substantially from the representation of others in that category" (Skiba et al, 2008, p. 266). Therefore, disproportionality can be either the over-representation or under-representation of a group in special education or a specific disability category. Two different aspects may be assessed when measuring disproportionality; the extent to which a group is differentially represented in a category compared to its proportion in the general population or the extent to which a group is differentially found eligible for special education services compared to that of other groups (Skiba et al, 2008).

The disproportionate representation of minority students in special education has been widely documented. In fact, monitoring requirements have been added to the

Individuals with Disabilities Education Act in order to assess the extent of disproportionality (Albrecht, Skiba, Losen, Chung, & Middelberg, 2012). Over-identification for special education placement can result in stigmatization, lowered expectations, reduced instruction, exclusion from the educational and social curricula of general education, and withdrawal from school (Cartledge, 2005; Reschly, 1996). Also, compared to similarly identified Caucasian peers, culturally and linguistically diverse students placed in special education experience less positive long-term outcomes in terms of enrollment in post-secondary education, employment, independent living, and incarceration (Affleck, Egar, Levine & Kortering, 1990).

The disproportionate representation of African American students is of particular concern as they are the most overrepresented group in special education in nearly every state (Parrish, 2002). Data collected by the Office of Special Education Programs (OSEP) and the Office for Civil Rights (OCR) on enrollment of students in special education programs broken down by racial/ethnic group indicated the following special education identification rates: 5% Asian/Pacific Islander, 11% Hispanic, 12% Caucasian, 13% American Indian, and 14% African American (Donovan & Cross, 2002). These statistics indicate that, compared to percentages in the general population, a higher percentage of African Americans are identified as i need of special education services than any other racial/ethnic group. Although African Americans have the greatest representation in all disability categories when compared to other races/ethnicities, the disproportionality is even more pronounced in the high-incidence categories of eligibility including learning disabilities, emotionally disabled, mild intellectual disability, and speech and language disorders. African American overrepresentation seems to be the most pronounced in the

high-incidence category of Intellectual Disability, with African American students more than twice as likely as Caucasian students to be labeled as such nationally (Cartledge & Dukes, 2008). In 2002, the general student population consisted of 17% African American students, but special education programs for Intellectually Disabled students consisted of 33% African American students (Donovan & Cross, 2002). The figures of African American overrepresentation in high-incidence categories do vary greatly according to region with the tendency for more pronounced overrepresentation in areas where the overall African American population is lower or in more affluent areas (Cartledge & Dukes, 2008).

Further, once identified as special education students, African Americans are at greater risk for more restrictive special education placements and are less likely to be provided access to the general education curriculum and environment in comparison to Caucasian peers (Skiba, Poloni-Staudinger, Gallini, Simmons, & Feggins-Aziz, 2006). So, not only do African American students have a greater representation in every disability area, but they are also found disproportionately in the most restrictive settings for every disability category. For example, Office of Civil Right data from 1998 indicates that 37% of African American special education students were served in an inclusive setting while 55% of Caucasian special education students were served in an inclusive setting. Thirty-three percent of African American special education students were served in a self-contain setting while 16% of Caucasian students were served in a self-contained setting (Fierros & Conroy, 2002). Another example of African American students' heightened risk of more restrictive placements, based on data from the state of Indiana, was provided by Skiba et al., 2006 and focused specifically on the high-incidence

categories of special education. In comparison to Caucasian students, African American emotionally disabled students were 1.2 times more likely to be served in a self-contained setting, African American mildly intellectually disabled students were 1.5 times more likely to be served in a self-contained setting, and African American learning disabled students were 3.2 times likely to be served in a self-contained setting (Skiba et al., 2006). Beyond the issue of restrictiveness, Cartledge and Dukes (2008) note that African American emotionally disabled students receive fewer services to address their needs, such as counseling, and are more frequently referred to the juvenile justice system when compared to Caucasian emotionally disabled students.

Purpose of the Current Study

"Screening for early literacy deficits is useful to the extent that the measures are accurate, sensitive to instructional needs, responsive to the effects of interventions, valid as predictors of later reading outcomes, and fair to all groups for whom inferences will be made " (Betts, et. al., 2008, p. 556). Given that measures free of predictive bias are essential to the effective use of assessment results in decision making, the examination of bias in predictive validity remains relatively uncommon. Of the information that is available, results are relatively inconsistent.

The purpose of the current study is to lend clarity to the current body of research on predictive bias in oral reading fluency through an investigation of racial, gender, and socioeconomic bias in DIBELS ORF probes for second through fifth grade African American and Caucasian students. Specifically, the difference in regression intercepts and slopes for Caucasian and African American second through fifth graders will be examined for evidence of predictive bias. Measures of Academic Progress (MAP), a

standardized, individually administered measure of reading, will be used as the criterion measure.

Research Questions

The following research questions will be addressed:

Research question 1. What are the predictive and concurrent relationships between DIBELS Oral Reading Fluency performance and reading comprehension scores on Measures of Academic Progress?

Hypothesis 1. It is expected that significant positive correlations exist between fall DIBELS Oral Reading Fluency performance and spring reading comprehension scores on Measures of Academic Progress and between spring DIBELS Oral Reading Fluency performance and spring reading comprehension scores on Measures of Academic Progress.

Research Question 2. Among fall DIBELS Oral Reading Fluency performance, spring DIBELS Oral Reading Fluency performance, race, gender, and socioeconomic status, what is the best variable, or combination of variables, in predicting spring reading comprehension scores on Measures of Academic Progress?

Hypothesis 2. It is expected that a combination of fall DIBELS Oral Reading Fluency performance, spring DIBELS Oral Reading Fluency performance, race, gender, and socioeconomic status will provide the strongest predictive utility in predicting spring reading comprehension scores on Measures of Academic Progress.

Research Question 3. Do DIBELS Oral Reading scores from fall and spring differentially predict reading comprehension scores on Measures of Academic Progress

across race (African American, Caucasian), gender (male, female), and socioeconomic group (free lunch, reduced-cost lunch, full-pay lunch)?

Hypothesis 3. It is expected that regression equations will differ significantly in slope, intercept, or both for the prediction of reading comprehension scores on Measures of Academic Progress by DIBELS Oral Reading Fluency scores across race (African American), gender (male, female), and socioeconomic group (free lunch, reduced-cost lunch, full-pay lunch).

Chapter 2

Method

Participants

Participants were 834 second through fifth grade students enrolled in three elementary schools in a school district located in a Southeastern US state. The district serves approximately 37,000 students in 51 schools. District enrollment consists of approximately 70% Caucasian students, 21.1% African American students, 7.4% Hispanic students, and 1.4% of other ethnicities. Sixty percent of students in the district receive free or reduced-cost lunch.

The current study is an analysis of predictive bias among Caucasian and African American students, therefore students of Hispanic, Asian, and other ethnicities were excluded ($N = 52$). Additionally, students who did not have complete test scores from all required points in time were excluded. The final sample consisted of all Caucasian ($n = 593$) and African American ($n = 241$) second through fifth grade students enrolled in the three elementary schools for whom complete test data were available. Demographic data, including race, gender, free, reduced, or full-pay lunch status, and special education status was obtained from district records at the time of the norming project.

The ethnic distribution of the final sample was approximately 33% African American and 67% Caucasian. The representation of African American students in the final sample was slightly higher than that of the total district enrollment due to the exclusion of other ethnicities from the study sample. Prior to their exclusion of other ethnicities, the sample aligned closely with the school district's demographic characteristics. Approximately 50% of the participants were female and 50% were male.

Lunch status was used as a proxy for socioeconomic status. Fifty-seven percent of the final sample was eligible for free or reduced-cost lunch. Approximately 16% of students in the sample received special education services. See Table A1 for further information regarding demographic information by grade.

Instruments

DIBELS Oral Reading Fluency. Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency (DIBELS ORF) is a standardized, individually administered measure of speed and accuracy in reading connected text for students in grades one through six (Good & Kaminski, 2002). All DIBELS ORF passages for a specific grade level are designed to match the end of year goal level of reading for that grade (Good & Kaminski, 2002). DIBELS ORF includes benchmark passages for screening purposes and 20 additional passages for progress monitoring purposes. Benchmark passages are administered three times throughout the school year (fall, winter, and spring). Students are required to read aloud a brief passage for one minute. The score for the passage is the number of words read correctly in one minute. Substitutions, omissions, and hesitations of more than three seconds are counted as errors. At each benchmark administration, three passages are administered. The benchmark score is the median of the three passage scores. For the purpose of the current study only DIBELS ORF benchmark scores were analyzed.

Many researchers have confirmed the technical adequacy of DIBELS ORF. Test-retest reliabilities for elementary students were found to range from .92 to .97; however, information regarding sample demographics was not reported (Tindal, Marston, & Deno, 1983). Test-retest reliability from the spring of first grade to the spring of second grade

was found to be .82 among a sample of 342 students of which 90% were Caucasian (Good, Simmons, & Kame'enui, 2001). Alternate form reliabilities were found to range from .87 to .93 among a sample of 134 second grade students (Francis et al., 2008). In a synthesis of psychometric evidence for DIBELS measures, Goffreda and DiPerna (2010) noted that DIBELS ORF is a reliable measure of reading performance for screening and group decision-making purposes according to measures of test-retest reliability and alternate form reliability.

Concurrent validity, as evidenced by seven peer-reviewed journal articles, two dissertations, and five technical reports reviewed in a recent empirical review of psychometric evidence for DIBELS, ranged from moderate to high among sample sizes ranging from 134 first graders to 35,207 third graders (Goffreda & DiPerna, 2010). Predictive validity coefficients for DIBELS ORF and statewide standardized achievement measures also ranged from moderate to high (Goffreda & DiPerna, 2010). Among 1,518 first grade students, 92% of which were African American, Reidel and Samuels (2007) reported a predictive validity coefficient of .69 with the TerraNova CAT Reading test. Roehrig, Petscher, Nettles, Hudson, and Torgesen (2008) reported predictive validity coefficients ranging from .66 to .68 for the Florida Comprehensive Assessment Test and .68 to .69 for the Stanford Achievement Test among a diverse sample of 35,207 third grade students. In a study of 2,588 first grade students, Schilling, Carlisle, Scott, and Zeng (2007) found a predictive validity coefficient of .69 for the Iowa Test of Basic Skills Reading Composite.

Measures of Academic Progress. Measures of Academic Progress (MAP), published by Northwest Evaluation Association (NWEA), is a computer-adapted test that

measures achievement in reading, mathematics, language, and science for students in grades two through ten (NWEA, 2003). MAP is administered three times throughout the school year; September, January, and April. For the purpose of this study, only MAP Reading scores were analyzed. All MAP test items are multiple choice; Reading items have four answer options. The Reading portion of MAP consists of four subareas; Word Meaning, Literal Comprehension, Interpretive Comprehension, and Evaluative Comprehension. The reading portion of MAP measures reading comprehension ranging from the single word level to comprehension of full text. Word Meaning items measure a student's word recognition and vocabulary skills. Literal Comprehension items measure a student's ability to recall, identify, classify, and sequence a variety of written material. Interpretive Comprehension items measure a student's ability to make predictions and draw inferences from written material. Evaluative Comprehension items measure a student's ability to understand fact, opinion, bias, assumption, and elements of persuasion: students are required to compare works, evaluate conclusions, and apply what was read.

Because of the computer-adaptive nature of MAP, each student receives a set of items optimal for their individual ability level (NWEA, 2003). The difficulty level of the first item presented is based on the examinee's previous MAP performance. If no previous MAP information is available, the first item presented is of average difficulty for the examinee's grade level. Following each item, the examinee's ability estimate is re-calculated and successive items are presented that match that ability estimate. MAP Reading tests are created from a pool of 1,200 items per grade level and are aligned with

state curriculum standards. A minimum of seven items per reading subarea are presented for each examinee.

MAP scores are reported in Rasch Units, or RIT scores. A RIT score is reported for each of the four achievement areas; reading, mathematics, language, and science. Scores are reported on a scale ranging from approximately 140 to 300. The MAP testing model is a one-parameter item response theory (IRT) model which places items and examinees on the same scale; therefore, MAP is useful for measuring growth across the school year, as well as growth across multiple grade levels (NWEA, 2003).

The technical manual for use with Measures of Academic Progress and Achievement Level Tests (NWEA, 2003) presents information regarding reliability and validity for MAP. Because MAP is administered multiple times throughout the year, test-retest reliability was calculated as the correlation between pre-instruction and post-instruction scores for the same student. Stability estimates ranged from .77 to .94 across grades two through ten. Standard errors of measurement are low across the RIT scale according to the technical manual (NWEA, 2003); however, no further information regarding standard errors of measure was provided.

Criterion-related and concurrent validity evidence was presented in the technical manual; however, correlations were conducted between NWEA's Achievement Level Tests (ALT) and other measures, not the computer-adaptive MAP tests. The ALT is a paper and pencil version of MAP with items drawn from the same bank as MAP items. In 2001, NWEA conducted a study of over 1,500 students who took both MAP and ALT. ALT tests were administered during the spring and MAP tests were administered the following fall. The validity coefficient for reading was .83: NWEA concluded that scores

from MAP and ALT are very closely correlated with ALT scores (NWEA, 2003). Further information regarding sample demographics for the validity study was not reported. Validity coefficients between 1999 ALT reading, mathematics, and language scores and Iowa Test of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2001) scores were calculated using a pool of 1,400 examinees in grades three, five, and nine: estimates ranged from .77 to .84. In 2001, validity coefficients between ALT reading, mathematics, and language scores and Stanford Achievement Tests, Ninth Edition (SAT-9; Harcourt Educational Measurement, 1996) scores were calculated using a pool of 4,000 second graders and 7,999 ninth graders: estimates ranged from .78 to .88. No other descriptive information regarding the samples was provided in the technical manual.

Procedure

This study is an analysis of existing data. Data used for the current study was collected by the participating school district during the 2008-2009 academic year. The dataset consisted of student fall and spring DIBELS ORF scores obtained from results of a local norming project and spring reading scores from an assessment administered district-wide three times per academic year. Three elementary schools were selected by the district for participation in the norming project due to their alignment with the school district's demographic characteristics. DIBELS ORF measures were administered to first through fifth grade students at select elementary schools as part of a local norming project. A group of trained assessors was utilized for DIBELS ORF data collection including school psychologists and academic interventionists who administer DIBELS measures on a regular basis as part of their job description. All assessors participating in

data collection for the norming project attended an eight hour DIBELS ORF administration training session prior to data collection.

DIBELS ORF measures, along with other curriculum-based measures included in the norming project, were administered at three time points throughout the 2008-2009 academic year: September, January, and April. At each administration, three DIBELS ORF measures were administered to each student: the median of the three spring scores was used in analysis. Use of the median DIBELS ORF score is a recommended best practice for use with DIBELS assessments (Good, Gruba, & Kaminski, 2002). Additionally, use of the median score controls for variance that may be caused by an extreme score (Hintze et al., 2002). The curriculum-based measures, which included DIBELS ORF, were administered in short ten to fifteen minute sessions. The three grade level DIBELS ORF passages were administered in the same sequential order to all students.

MAP assessments were administered to all students in grades two through eight during the fall, winter, and spring in partial fulfillment of the mandate for an accountability system by the South Carolina Education Act of 1998 (South Carolina Education Accountability Act, 1998, Section 59-18-300). MAP assessments were administered by trained assessors following NWEA's standardized administration procedures (NWEA, 2003). Only MAP Reading scores from the spring were analyzed in this study.

Analyses

Reynolds and Carson (2005) advocate that any investigation of predictive bias begin with an omnibus test, then follow up tests for specific group differences, and

identification of specific group differences in slope, *y*-intercept, or both. "Potthoff (1978) provides an efficient and parsimonious regression bias procedure that allows both simultaneous and separate tests of regression slopes and intercepts across groups" (Watkins & Hetrick, 1999, p. 710). Because it allows for a single, simultaneous test of equivalence of slope and *y*-intercept differences, unlike alternative methods, Potthoff's procedure reduces the probability of Type I errors (Konold & Canivez, 2010). Researchers have consistently demonstrated a preference for the use of Potthoff's procedure to examine bias in predictive validity among diverse subgroups (Bossard, Reynolds & Gutkin, 1980; Canivez & Konold, 2001; Glutting, 1986; Glutting, Oakland & Konold, 1994; Konold & Canivez, 2010; Naglieri & Hill, 1986; Reynolds & Hartlage, 1979; Shields, Konold & Glutting, 2004; Weiss & Prifitera, 1995). To investigate the presence of bias of DIBELS ORF in predicting MAP reading comprehension scores across race, gender, and socioeconomic status, Potthoff's (1978) procedure was used.

To examine the research question, median spring DIBELS ORF scores were used to predict spring MAP reading comprehension scores. Because DIBELS ORF and MAP items differ at each grade level, a comparison across grades is not appropriate; therefore, separate analyses were conducted for each grade (Kranzler et al., 1999). Equality of slopes and *y*-intercepts were examined across race, gender, and socioeconomic status for each grade level. Due to the categorical nature of the race, gender, and socioeconomic status variables, these variables were transformed into dummy coded variables to allow their inclusion in multiple regression analyses. The omnibus simultaneous *F* test was first conducted to determine the presence of bias. Following the omnibus test, further

examination of statistically significant group comparisons was conducted in order to determine source of bias: slope, *y*-intercept, or both.

Chapter 3

Results

Descriptive Statistics

Means and standard deviations for fall and spring DIBELS ORF and spring MAP Reading scores are reported for each total grade level sample, as well as for demographic groups. See Tables A2 through A5 for descriptive statistics for grades 2 through 5, respectively. An inspection of mean scores for the three measures revealed some patterns. As expected, means for fall DIBELS ORF scores are generally less than means for spring DIBELS ORF scores with the exception of African American fifth graders whose fall DIBELS ORF mean of 98.23 is greater than the spring DIBELS ORF mean 97.77, although this is not statistically significant. For all three measures across all grade levels, African Americans earned lower mean scores compared to Caucasians. Results of *t* Tests indicate that 10 of the 12 group mean comparisons between Caucasians and African American students were statistically significant (see Table A6). Additionally, the full-pay lunch group tended to earn higher mean scores on all three measures in comparison to the reduced lunch group, and the reduced lunch group tended to earn higher mean scores on all three measures in comparison to the free lunch group. Exceptions to this pattern of score attenuation included the following: grade 2 fall DIBELS ORF (free lunch group mean = 61.49; reduced lunch group mean = 55.82), grade 5 fall DIBELS ORF (free lunch group mean = 111.06; reduced lunch group mean = 107.44), and grade 5 MAP (free lunch group mean = 214.59; reduced lunch group mean = 211.11). Results of *t* Tests indicate that 15 of the 36 group mean comparisons between lunch status groups were

statistically significant (see Table A6). For gender, the only statistically significant group mean comparison was for 4th grade spring DIBELS ORF scores (see Table A6).

Research Question #1

The first research question was: What are the predictive and concurrent relationships between DIBELS Oral Reading Fluency performance and reading comprehension scores on Measures of Academic Progress?

For each grade level, correlations among all measures were calculated and are presented in correlation matrixes (see Tables A7 through A10). Next, correlation analyses were calculated to determine the predictive and concurrent relationships between DIBELS ORF and MAP reading comprehension scores for each grade level total sample and grade level demographic groups. Correlations for grades 2, 3, 4, and 5 are presented in Tables A11, A12, A13, and A14, respectively. The concurrent administrations between DIBELS ORF (spring administration) and reading comprehension scores on MAP yielded the strongest correlations for all grade levels. For these concurrent administrations, correlation coefficients for all demographic groups, as well as each total grade level were statistically significant ($p < .01$) and were moderate to large, ranging from .56 to .81. However, correlations between fall DIBELS ORF administration and MAP reading comprehension scores were generally weak and some were negative.
For grade 2, correlations for fall DIBELS ORF and MAP reading comprehension scores ranged from -.21 to .22. The only statistically significant correlation between fall DIBELS ORF and MAP reading comprehension scores for grade 2 was for the full-pay lunch group ($p < .05$). Correlations for spring DIBELS ORF and MAP reading

comprehension scores across all demographic groups were statistically significant ($p <$ .01) and ranged from .65 to .81.

For grade 3, correlations for fall DIBELS ORF and MAP reading comprehension scores ranged from -.19 to .15. None of the correlations between fall DIBELS ORF and MAP reading comprehension scores were statistically significant. Correlations for spring DIBELS ORF and MAP reading comprehension scores across all demographic groups were statistically significant ($p < .01$) and ranged from .62 to .74.

For grade 4, correlations for fall DIBELS ORF and MAP reading comprehension scores ranged from -.09 to .29. Correlations between fall DIBELS ORF and MAP reading comprehension scores for the grade 3 total sample, Caucasians, males, and the full-pay lunch group were statistically significant ($p < .01$). Correlations for spring DIBELS ORF and MAP reading comprehension scores across all demographic groups were statistically significant ($p < .05$) and ranged from .56 to .71.

For grade 5, correlations for fall DIBELS ORF and MAP reading comprehension scores ranged from -.04 to .23. Correlations between fall DIBELS ORF and MAP reading comprehension scores for the grade 3 total sample and for females were statistically significant ($p < .05$). Correlations for spring DIBELS ORF and MAP reading comprehension scores across all demographic groups were statistically significant ($p <$ .01) and ranged from .58 to .81.

Research Question #2

The second research question was: Among fall DIBELS Oral Reading Fluency performance, spring DIBELS Oral Reading Fluency performance, race, gender, and

socioeconomic status, what is the best variable, or combination of variables, in predicting spring reading comprehension scores on Measures of Academic Progress?

Stepwise multiple regression analyses were conducted for each grade level to examine the best variable, or combination of variables, in the prediction of MAP reading comprehension scores. For each grade level, fall DIBELS ORF, spring DIBELS ORF, race, gender, and socioeconomic status were entered into the regression equation as predictors and the criterion variable was MAP reading comprehension scores. Models differed across the 4 grade levels, however, spring DIBELS ORF scores were included, whether alone or in combination with other predictors, in the best model for all grades.

Stepwise regression results for grade 2 are presented in Table A15. The final model included two predictor variables, spring DIBELS ORF and race. Model 1, which included only spring DIBELS ORF, accounted for 58 percent of the variance in MAP reading comprehension scores, R^2 adj. = .58, $\Delta F(1, 238) = 327.77$, $p < .01$. The inclusion of race into model 2 resulted in an additional 3 percent of the variance of MAP reading comprehension scores being explained, $\Delta F(2, 237) = 17.95$, $p < .01$. Beta weights were statistically significant for both predictors in the final model, as indicated by *t*-statistics (see Table A16). These results indicated that, for grade 2, the best set of predictors in predicting MAP reading comprehension scores was spring DIBELS ORF and race.

Stepwise regression results for grade 3 are presented in Table A17. The model included one predictor variable, spring DIBELS ORF. This model accounted for 48 percent of the variance in MAP reading comprehension scores, R^2 adj. = .48, $\Delta F(1, 215) = 198.10$, $p < .01$. The Beta weight is statistically significant for spring DIBELS ORF in the model, as indicated by *t*-statistics (see Table A18). These results indicated that, for

grade 3, the best predictor in predicting MAP reading comprehension scores was spring DIBELS ORF.

Stepwise regression results for grade 4 are presented in Table A19. The final model included two predictor variables, spring DIBELS ORF and lunch status. Model 1, which included only spring DIBELS ORF, accounted for 47 percent of the variance in MAP reading comprehension scores, R^2 adj. = .47, ΔF (1, 213) = 190.59, $p < .01$. The inclusion of lunch status into model 2 resulted in the addition of only 2 percent of the variance of MAP reading comprehension scores being explained, ΔF (1, 212) = 8.92, $p < .01$. Beta weights are statistically significant for both predictors in the final model, as indicated by *t*-statistics (see Table A20). These results indicated that, for grade 4, the best set of predictors in predicting MAP reading comprehension scores was spring DIBELS ORF and lunch status.

Stepwise regression results for grade 5 are presented in Table A21. The final model included three predictor variables, spring DIBELS ORF, race, and lunch status. Model 1, which included only spring DIBELS ORF, accounted for 49 percent of the variance in MAP reading comprehension scores, R^2 adj. = .49, ΔF (1, 160) = 157.32, $p < .01$. The inclusion of race into model 2 resulted in an additional 5 percent of the variance of MAP reading comprehension scores being explained, ΔF (1, 159) = 15.82, $p < .01$. The inclusion of lunch status in model 3 resulted in an additional 2 percent of the variance of MAP reading comprehension scores being explained, ΔF (1, 158) = 6.95, $p < .01$. Beta weights were statistically significant for the three predictors in the final model, as indicated by *t*-statistics (see Table A22). These results indicate that, for grade 5, the

best set of predictors in predicting MAP reading comprehension scores was spring DIBELS ORF, race, and lunch status.

Research Question #3

The third research question was: Do DIBELS Oral Reading scores from fall and spring differentially predict reading comprehension scores on Measures of Academic Progress across race (African American, Caucasian), gender (male, female), and socioeconomic group (free lunch, reduced-cost lunch, full-pay lunch)?

For each grade level, six simultaneous demographic group comparisons were conducted via Potthoff's procedure to determine whether DIBELS ORF scores from fall and spring differentially predict reading comprehension scores on Measures of Academic Progress across race (African American, Caucasian), gender (male, female), and socioeconomic group (free lunch, reduced-cost lunch, full-pay lunch). Tables A23 through A26 present *F* values, degrees of freedom, and corresponding *p* values for all simultaneous contrasts of slope and intercept differences between demographic groups for grades 2 through 5, respectively. Nine of the 24 simultaneous contrasts were statistically significant ($p < .01$). Seven of the nine statistically significant contrasts were related to race, one was related to gender, and one was related to lunch status. Follow-up evaluations of the statistically significant omnibus demographic group comparisons were conducted to determine if demographic groups differed significantly in slope, intercept, or both.

Race. Simultaneous slope and intercept Potthoff comparisons between African Americans and Caucasians revealed differential prediction of MAP reading comprehension scores from fall DIBELS ORF for all grades 2 through 5. However,

follow-up tests of slope and intercept revealed no significant differences. Simultaneous slope and intercept Potthoff comparisons between racial groups revealed differential prediction of MAP reading comprehension score from spring DIBELS ORF for grades 2, 4, and 5. For grade 2, no significant differences in slope or intercept were found. For grade 4, racial groups differed significantly in intercept. For grade 5, racial groups differed significantly in both slope and intercept.

Gender. Simultaneous slope and intercept Potthoff comparisons between gender groups did not reveal differential prediction of MAP reading comprehension scores from fall DIBELS ORF for any grade. Simultaneous slope and intercept Potthoff comparisons between gender groups revealed differential prediction of MAP reading comprehension scores from spring DIBELS ORF for grade 2 only; gender groups differed significantly in both slope and intercept.

Lunch Status. Simultaneous slope and intercept Potthoff comparisons between lunch status groups did not reveal differential prediction of MAP reading comprehension scores from fall DIBELS ORF for any grade. Simultaneous slope and intercept Potthoff comparisons between lunch status groups revealed differential prediction of MAP reading comprehension scores from spring DIBELS ORF for grade 3 only; lunch status groups differed significantly in intercept only.

Chapter 4

Discussion

Research Summary

IDEIA (2004) provisions allowing educational agencies to use Response to Intervention (RTI) in determining whether a child has a specific learning disability coupled with implementation of the No Child Left Behind Act in 2001 which placed a focus on large-scale testing and accountability (NCLB, 2001) resulted in the increasing use of CBM as a standardized measurement tool for understanding students' progress towards and achievement of state standards, particularly in reading through the use of R-CBM or oral reading fluency measures such as DIBELS ORF. The unique features of R-CBM, including its psychometric properties, ability to function as a general outcome measure, and the ease of administration, time efficiency, low cost, and frequency with which the measures may be given, has led to widespread use in U.S. schools. These same properties make R-CBM's, such as DIBELS ORF, worthy of analysis as a direct measure of reading fluency and as a correlate to reading comprehension and general reading proficiency (Reschly, et al., 2009).

Extensive evidence of oral reading fluency's predictability of reading comprehension exists, but little research on the differential prediction, or predictive bias, of racial, gender, or socioeconomic subgroups is available. Since much of the value of a screening measure is determined by its ability to predict future outcomes on a criterion measure, the extent to which the inferences of future performance hold true for all subpopulations of interest is an essential area of investigation (Betts et al., 2008). The examination of bias in predictive validity of oral reading fluency measures remains

relatively uncommon. A handful of available studies directly examine predictive bias in R-CBM; however, no clear pattern of differential prediction has been consistent across race, gender, or grade level (Hosp et al., 2011).

Because no pattern has been established, caution in the use of oral reading fluency probes with diverse students is warranted. The continuation of a rigorous examination of possible bias with racial, gender, and socioeconomic groups through diverse psychometric techniques is recommended (Reynolds & Ramsay, 2003). Because CBMs are widely used for both identification for remediation programs in regular education and the identification of students with specific learning disabilities, possible predictive bias in CBMs may contribute significantly to disproportionality in special education and to inequitable provision of remediation programs provided through general education. The disproportionate representation of African American students is of particular concern as they are the most overrepresented group in special education in nearly every state and they are also at greater risk for more restrictive special education placements for the same disability compared to Caucasian peers.

The purpose of the current study was to lend clarity to the current body of research on predictive bias in oral reading fluency through an investigation of racial, gender, and socioeconomic bias in DIBELS Oral Reading Fluency probes for second through fifth grade African American and Caucasian students. Before an analysis of predictive bias was conducted, the strength of relationship and the nature of the predictive relationship between DIBELS ORF and MAP reading comprehension scores in this sample of students were explored. To accomplish these purposes, the following research questions were addressed: (1) What are the predictive and concurrent relationships

between DIBELS Oral Reading Fluency performance and reading comprehension scores on Measures of Academic Progress? (2) Among fall DIBELS Oral Reading Fluency performance, spring DIBELS Oral Reading Fluency performance, race, gender, and socioeconomic status, what is the best variable, or combination of variables, in predicting spring reading comprehension scores on Measures of Academic Progress? and (3) Do DIBELS Oral Reading scores from fall and spring differentially predict reading comprehension scores on Measures of Academic Progress across race (African American, Caucasian), gender (male, female), and socioeconomic group (free lunch, reduced-cost lunch, full-pay lunch)?

Research Question #1

To answer the first research question, correlation analyses were calculated to evaluate the strength of concurrent and predictive relationship between DIBELS ORF and MAP reading comprehension scores. As expected, the concurrent administrations between DIBELS ORF (spring administration) and reading comprehension scores on MAP yielded the strongest correlations for all grade levels. For these concurrent administrations, correlation coefficients for all demographic groups, as well as each total grade level were statistically significant ($p < .01$) and were moderate to large, ranging from .56 to .81. This finding confirms previous research indicating a positive relationship between reading fluency and reading comprehension (Meyer & Felton, 1999; Reschly et al., 2009; Shinn et al.,1992). However, the same did not hold true for the predictive relationship between fall DIBELS ORF administration and MAP reading comprehension scores. These correlations were generally weak and some were even negative. Previous research has indicated stronger relationships between measures of reading fluency and

measures of reading comprehension when time intervals between administrations are shorter (Baker et al., 2008; Roehrig et al., 2008); however, data for the current study was gathered within the same school year making the lack of significant correlation for many grades and demographic groups a somewhat surprising finding. Among this sample of African American and Caucasian students, oral reading fluency, as measured by DIBELS ORF, is generally not related to future reading comprehension performance, as measured by MAP. In sum, the first research hypothesis was partially supported: significant positive correlations were found between spring DIBELS ORF performance and spring reading comprehension scores on Measures of Academic Progress, but correlations between fall DIBELS Oral Reading Fluency performance and spring reading comprehension scores on Measures of Academic Progress were, for the most part, weak. Although the concurrent relationship between DIBELS ORF and MAP reading comprehension is moderate to strong, the predictive relationship from fall to spring of the same academic year is weak suggesting that DIBELS ORF is not related to future reading comprehension performance in this sample of students.

Research Question #2

To answer the second research question, stepwise multiple regression analyses were conducted for each grade level to examine the best variable, or combination of variables in predicting MAP reading comprehension scores. For each grade level, the best prediction model differed; however, spring DIBELS ORF scores were included, whether alone or in combination with other predictors, in the best prediction model for all grades. For grade 2, the best set of predictors in predicting MAP reading comprehension scores was spring DIBELS ORF and race. Spring DIBELS ORF alone accounted for 58% of the

explained variance in MAP reading comprehension scores, while race accounted for only an additional 3 percent. For grade 3, spring DIBELS ORF alone was the best predictor in predicting MAP reading comprehension scores and accounted for 48% of the explained variance in MAP reading comprehension scores. For grade 4, the best set of predictors in predicting MAP reading comprehension scores was spring DIBELS ORF and lunch status. Alone, spring DIBELS ORF accounted for 47% of the explained variance in MAP reading comprehension scores, while lunch status accounted for only an additional 2%. For grade 5, the best set of predictors in predicting MAP reading comprehension scores was spring DIBELS ORF, race, and lunch status. Alone, spring DIBELS ORF accounted for 49% of the explained variance in MAP reading comprehension scores, while race accounted for only an additional 5% and lunch status accounted for only an additional 2%. The second hypothesis was partially supported. Spring DIBELS ORF scores did contribute significantly to the best prediction model across all four grades. Fall DIBELS ORF scores; however, were not useful in predicting MAP reading comprehension. Further, there is no consistent pattern of demographic variables contributing to the prediction of MAP reading comprehension performance. The results of stepwise regression analyses, coupled with the correlation results, emphasize the need for caution in using fall DIBELS ORF scores in predicting future success on high stakes testing or general reading comprehension performance. DIBELS ORF may be a poor predictor of high stakes test performance or reading comprehension in some students, as demonstrated in this sample of African American and Caucasian students. Fall DIBELS ORF may be less predictive of MAP reading comprehension performance for some students due to the implementation of interventions throughout the school year. Theoretically, in an RTI

model, students whose fall DIBELS ORF scores are below a predetermined cut-off would receive intervention as a supplement to, or replacement for, the core reading curriculum. If these interventions are successful and student's reading ability truly improves as a result, spring DIBELS ORF scores and spring MAP reading comprehension scores may be a more accurate reflection of their reading performance at that point in time rather than the prediction based on the fall score that was achieved prior to intervention.

Research Question #3

To answer the third research question, six simultaneous demographic group comparisons for each grade were conducted via Potthoff's procedure to determine if DIBELS ORF scores from fall and spring differentially predict reading comprehension scores on Measures of Academic Progress across race (African American, Caucasian), gender (male, female), and socioeconomic group (free lunch, reduced-cost lunch, full-pay lunch). Across the four grade levels, eight contrasts were conducted to examine racial bias; seven of these omnibus Potthoff analyses yielded a significant effect for race for the prediction of MAP reading comprehension scores. However, four of these contrasts were between fall DIBELS ORF and MAP reading comprehension (grades 2, 3, 4, and 5) and none resulted in significant slope or intercept differences indicating that no clinically or practically meaningful differences between groups were observed and the common regression line is appropriate for prediction for both Caucasian and African American students.

The remaining three significant racial contrasts were between spring DIBELS ORF and MAP reading comprehension scores for grades 2, 4, and 5. Follow-up comparisons of slope and intercepts revealed a significant intercept difference for grades

2 and 4, and significant slope and intercept differences for grade 5. As previously stated (Reynolds, 1984), when significant intercept differences are found the group with the lower mean criterion score is over-predicted when the common regression line is used. African American students obtained noticeably lower mean MAP reading comprehension scores across all four grade levels. Therefore, these results suggest that African American students' MAP reading comprehension scores will be over-predicted in relation to the common regression line when spring DIBELS ORF scores are used as predictors.

As previously stated, predictive validity is defined as the effectiveness of a test in predicting an individual's performance in specified activities (Anastasi, 1988). Predictive bias occurs when one regression equation is incorrectly used for two or more groups. The third hypothesis stated that regression equations will differ across race, gender, and socioeconomic group. Alternatively stated, predictive bias will be found. This hypothesis was partially supported. Regression equations predicting reading comprehension scores on Measures of Academic Progress from spring DIBELS ORF for African Americans and Caucasians differed significantly in intercept for two of four grades analyzed and in both slope and intercept for one grade analyzed. In sum, racial bias in predicting MAP reading comprehension performance from spring DIBELS ORF was found. The use of a common regression line for both groups is not appropriate and leads to over-prediction of the performance of African American students.

In an RTI model, this over-prediction of African American students on the criterion measure results in the under-identification for compensatory programs due to the fact that DIBELS ORF over-predicts their actual reading comprehension skills. Conversely, for Caucasian students, DIBELS ORF could potentially over-identify the

need for compensatory remediation in reading because their performance on screening measures may underestimate their true reading comprehension abilities. Curriculum-based measures used for universal screening are often characterized by high rates of under- or over-identification which have been shown to differentially affect different subgroups of students (Cleary et al., 1975; Hosp et al., 2011). The results of this study indicate that African American students are likely to be under-identified for compensatory programs through regular education. This, in turn, may lead to their over-representation, or disproportionality, in special education as they are less likely to be provided needed regular education interventions, increasing their likelihood to be referred for special education services.

African American students are disproportionately overrepresented in special education in nearly every state compared to percentages in the general population and a higher percentage of African Americans are identified as in need of special education services than any other racial/ethnic group (Parrish, 2002). This disproportionality is even more pronounced in the high-incidence categories of eligibility, which includes SLD. As discussed previously, RTI models of SLD identification rely heavily on the use of CBM's. When these measures display bias against members of a certain group, as found in the current study, they are ultimately contributing to disproportionality in special education.

Limitations

The sample of participants for this study were selected from one school district in a southeastern state and consisted of Caucasian and African American students only. Although the differential prediction of DIBELS ORF among Caucasian and African

American students is of significant importance, the findings here may not extend to other races/ethnicities or represent the relationship between oral reading fluency and reading comprehension in the general population. Further research that extends to other geographic locations and other races/ethnicities is necessary to determine the generalization of findings.

Results of this study are limited to the specific assessments used to measure oral reading fluency and reading comprehension. Because a myriad of assessments measuring reading comprehension are available, the concurrent and predictive relationships of DIBELS ORF, as well as alternative measures of oral reading fluency such as the STEEP (Witt, 2007), with alternative measures of reading comprehension needs to be examined before conclusions regarding differential prediction among demographic groups can be established.

In this study, fall DIBELS ORF scores were found to have a much weaker relationship with MAP reading comprehension scores than spring DIBELS ORF scores. Several possible extraneous factors may have contributed to this difference in concurrent and predictive relationships. The effects of instruction and intervention were not accounted for in this study. Theoretically, students scoring lowest on the fall DIBELS ORF administration would have been provided with intensive research-based intervention, whereas those scoring higher on the fall DIBELS ORF administration would have received only the core reading instruction. Additionally, because this was an existing data set, students who had missing or incomplete data were not able to be analyzed and therefore, the nature of those participants is unknown.

Directions for Future Research

While the current study contributes to the body of literature regarding differential prediction of reading comprehension from DIBELS ORF, further investigation is needed. In general, CBMs are considered to be useful to the extent that they are accurate, sensitive to instructional needs, responsive to the effects of intervention, valid as predictors of later reading outcomes, and fair to all groups for whom inferences will be made (Betts, et. al., 2008). Given that measures free of bias are essential to the effective use of assessment results in decision making, the examination of bias in concurrent and predictive validity across all populations of interest are necessary. The research questions addressed here should be extended to all races, ethnicities and geographic locations where DIBELS ORF are utilized in predicting success on high stakes testing or assisting in special education eligibility decisions. Additionally, research on predictive bias should be extended to alternative measures of both reading fluency and reading comprehension.

Researchers may consider the examination of alternative measures used in combination with DIBELS ORF to enhance the validity in predicting reading comprehension. Of particular interest is the use of multiple measures in predicting reading comprehension for minority groups whose reading comprehension is typically over-predicted in relation to the common regression line (Hintze et al., 2002). If such measures are found to increase the predictive validity among minority students, education agencies may be able to enhance their targeted intervention and instruction to those truly in need.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

Affleck, J. Q., Edgar, E., Levine, P., & Kortering, L. (1990). Postschool status of students classified as mildly mentally retarded, learning disabled, or nonhandicapped: Does it get better with time? *Education & Training in Mental Retardation, 25,* 315-324.

Albrecht, S. F., Skiba, R. J., Losen, D. J., Chung, C.-G., & Middelberg, L. (2012). Federal policy on disproportionality in special education: Is it moving us forward? *Journal of Disability Policy Studies, 23,* 14-25.

Anastasi, A. (1988). Testing the test: Interpreting results from multiscore batteries. *Journal of Counseling & Development, 64,* 84-86.

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame'enui, E. J., & Beck, C. T. (2008). Reading fluency as a predictor of reading proficiency in low-performing, high-poverty schools. *School Psychology Review, 37,* 18-37.

Betts, J., Reschly, A., Pickart, M., Heistad, D., Sheran, C., & Marston, D. (2008). An examination of predictive bias for second grade reading outcomes from measures of early literacy skills in kindergarten with respect to English-language learners and ethnic subgroups. *School Psychology Quarterly, 23,* 553-570.

Bossard, M.D., Reynolds, C. R., & Gutkin, T. B. (1980). A regression analysis of test bias on the Stanford-Binet Intelligence Scale for Black and White children referred for psychological services. *Journal of Clinical Child Psychology, 9,* 52-54.

Brown, R. T., Reynolds, C. R., & Whitaker, J. S. (1999). Bias in mental testing since bias in mental testing. *School Psychology Quarterly, 14,* 208-238.

Busch, T. W., & Reschly, A. L. (2007). Progress monitoring in reading. *Assessment for Effective Intervention, 32,* 223-230.

Canivez, G. L., & Konold, T. R. (2001). Assessing differential prediction bias in the Developing Cognitive Abilities Test across gender, race/ethnicity, and socioeconomic groups. *Educational and Psychological Measurement, 61,* 159-171.

Cartledge, G. (2005). Restrictiveness and race in special education: The failure to prevent or return. *Learning Disabilities: A Contemporary Journal, 3,* 27-32.

Christ, T. J., Burns, M. K., & Ysseldyke, J. E. (2005). Conceptual confusion within response to intervention vernacular: Clarifying meaningful differences. *Communique, 34*. Retrieved from http://www.nasponline.org/publications/cq/cq343rti.aspx

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1975). Educational use of tests with disadvantaged students. *American Psychologist, 30,* 15-41.

Daane, M.C., Campbell, J. R., Grigg, W. S., Goodman, M.J., & Oranje, A. (2005*). Fourth-grade students reading aloud: NAEP 2002 special study of oral reading (NCES 2006-469)*. Washington, DC: National Center for Educational Statistics, U.S. Department of Education, Institute of Education Sciences.

Deno, S. (1985).Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52,* 219-232.

Deno, S. L. (2003). Curriculum-based measures: Development and perspectives. *Assessment for Effective Intervention, 28,* 3-12.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37,* 184-192.

Deno, S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual*. Minneapolis, MN: Leadership Training Institute/Special Education, University of Minnesota.

Federal Register. (2006), Part II. 34 CFR Parts 300 and 301 Rules and Regulations, Volume 71, Number 156. Washington, DC: U.S. Department of Education.

Flanagan, D. P., Fiorello, C. A., & Ortiz, S. O. (2010). Enhancing practice through application of Cattell-Horn-Carroll theory and research: A "Third Method" approach to specific learning disability identification. *Psychology in the Schools, 47,* 739-760.

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*, 315-342.

Fuchs L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57*, 488-500.

Fuchs, L. S., & Deno, S. L. (1994). Must instructionally useful performance assessment be based in the curriculum? *Exceptional Children, 61,* 15-24.

Fuchs, L. S. & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21*, 45-58.

Fuchs, D. & Fuchs, L. S. (2005). Responsiveness-to-intervention: A blueprint for practitioners, policymakers, and parents. *Teaching Exceptional Children, 38,* 57-61.

Fuchs, D. & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly, 41,* 93-99.

Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research & Practice, 18,* 157-171.

Glutting, J. J. (1986). Potthoff bias analyses for K-ABC MPC and Nonverbal Scale IQs among Anglo, Black, and Puerto Rican kindergarten children. *Professional School Psychology, 1,* 225-234.

Glutting, J. J., Oakland, T., & Konold, T. R. (1994). Criterion-related bias with the Guide to the Assessment of Test-Session Behavior for the WISC-III and WIAT: Possible race, gender, and SES effects. *Journal of School Psychology, 32,* 355-369.

Goffreda, C. T., & DiPerna, J. C. (2010). An empirical review of psychometric evidence for the Dynamic Indicators of Basic Early Literacy Skills. *School Psychology Review, 39,* 463-483.

Good, R. H., Gruba, J., & Kaminski, R. A. (2002). Best practices in using dynamic indicators of basic early literacy skills (DIBELS) in an outcomes-driven model. In A. Thomas & J. Grimes (Eds.), *Best Practices in School Psychology IV* (pp. 699-720). Washington, DC: National Association of School Psychologists.

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for Development of Educational Achievement.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5,* 257-288.

Goodman, K. S. (Ed.). (2006). *The truth about DIBELS: What it is and what it does.* Portsmouth, NH: Heinemann.

Hale, J. B., Kaufman, A., Naglieri, J. A., & Kavale, K. A. (2006). Implementation of IDEA: Integrating response to intervention and cognitive assessment methods. *Psychology in the Schools, 43,* 753-770.

Harcourt Educational Measurement. (1996). Stanford Achievement Test Series, Ninth Ed. San Antonio, TX: Harcourt Educational Measurement.

Hasbrouck, J. & Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59,* 636-644.

Hintze, J. M., Callahan, J. E., Matthews, W. J. Williams, A. S., & Tobin, K. G. (2002). Oral reading fluency and prediction of reading comprehension in African American and Caucasian elementary school children. *School Psychology Review, 31,* 540-553.

Hintze, J. M., Christ, T. J., & Methe, S. A. (2006). Curriculum-based assessment. *Psychology in the Schools, 43,* 45-56.

Hosp, J. L. & Ardoin, S. (2008). Assessment for instructional planning. *Assessment for Effective Intervention, 33,* 69-77.

Hosp, J. L., & Reschly, D. J. (2003). Referral rates for intervention or assessment: A meta-analysis of racial differences. *Journal of Special Education, 37,* 67-80.

Hosp, J. L., Hosp, M. A., & Dole, J. K. (2011). Potential bias in predictive validity of universal screening measures across disaggregation subgroups. *School Psychology Review, 40,* 108-131.

Hixson, M. D. & McGlinchey, M. T. (2004). The relationship between race, income, and oral reading fluency and performance on two reading comprehension measures. *Journal of Psychoeducational Assessment, 22,* 351-364.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2001). Iowa Tests of Basic Skills (ITBS). Itasca, IL: Riverside Publishing.

Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C. § 1400 et seq. (2004).

Jensen, A.R. (1980). *Bias in Mental Testing*. New York: Free Press.

Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review, 37,* 374-390.

Klein. J. R. & Jimmerson, S. R. (2005). Examining ethnic, gender, language, and socionomic bias in oral reading fluency scores among Caucasian and Hispanic students. *School Psychology Quarterly, 20,* 23-50.

Konold, T. R., & Canivez, G. L. (2010). Differential relationships between WISC-IV and WIAT-II scales: An evaluation of potentially moderating child demographics. *Educational and Psychological Measurement, 70,* 613-627.

Kranzler, J. H., Miller, M. D., & Jordan, L. (1999). An examination of racial/ethnic and gender bias on curriculum-based measurement of reading. *School Psychology Quarterly, 14,* 327-342.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6,* 293-323.

Marston, D. (1989). Curriculum-based measurement: What is it and why do it? In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guilford Press.

Meyer, M. S., & Felton, R. H. (1999). Repeated reading to enhance fluency: Old approaches and new directions. *Annals of Dyslexia, 49,* 283-306.

Naglieri, J. A., & Hill, D. S. (1986). Comparison of WISC-R and K-ABC regression lines for academic prediction with Black and White children. *Journal of Clinical Child Psychology, 15,* 352-355.

National Center on Response to Intervention. (2010). Essential components of RTI: A closer look at response to intervention. Washington, DC: U.S. Department of Education, Office of Special Education Programs, National Center on Response to Intervention. Retrieved from http://www.rti4success.org/pdf/rtiessentialcomponents_042710.pdf

National Dissemination Center for Children with Disabilities (NICHCY) (2012). Response to Intervention (RTI). Retrieved from http://nichcy.org/schools-administrators/rti

No Child Left Behind Act of 2001, 20 U.S.C. 70 § 6301 et seq. (2002).

Northwest Evaluation Association NWEA (2003). Technical manual for use with Measures of Academic Progress. Portland, OR: Northwest Evaluation Association. Retrieved from http://www.apsrc.net/Images/Interior/nwea%20resources/nwea_technicalmanual.pdf

Parrish, T. (2002). Racial disparities in the identification, funding, and provision of special education. In D. J. Losen & G. Orfield (Eds.), *Racial inequity in special education* (pp. 15-37). Cambridge, MA: Harvard Education Press.

Pearce, L. R., & Gayle, R. (2009). Oral reading fluency as a predictor of reading comprehension with American Indian and White elementary students. *School Psychology Review, 38,* 419-427.

Potthoff. R. F. (1978). *Statistical aspects of the problem of biases in psychological tests* (Institute of Statistics Mimeo Series No. 479). Chapel Hill: University of North Carolina, Department of Statistics.

Reidel, B. W., & Samuels, S. J. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42,* 546-567.

Reschly, D. J. (1996). Identification and assessment of students with disabilities. *The Future of Children, 6,* 40-53.

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-Based Measurement Oral Reading as an indicator or reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47,* 427-469.

Reynolds, C. R. (1982). Construct and predictive bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 199-227). Baltimore: Johns Hopkins University Press.

Reynolds, C. R., & Hartlage, L. (1979). Comparison of WISC and WISC-R regression lines for academic prediction with Black and with White referred children. *Journal of Consulting and Clinical Psychology, 47,* 589-591.

Reynolds, C. R., Lowe, P. A., & Saenz, A. (1999). The problem of bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (549-595). New York: Wiley.

Reynolds, C. R. & Ramsay, M. C. (2003). Bias in psychological assessment: An empirical review and recommendations. *Handbook of psychology: Assessment Psychology, 10,* 67-93.

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS Oral Reading Fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology, 46,* 343-366.

Samuels, S. J. (1979). The method of repeated readings. *The Reading Teacher, 32,* 403-408.

Schilling, S. G., Carlisle, J. F., Scott, S. E., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *The Elementary School Journal, 107,* 429-448.

Shields, J., Konold, T. R., & Glutting, J. J. (2004). Validity of the Wide Range Intelligence Test: Differential effects across race/ethnicity, gender, and education level. *Journal of Psychoeducational Assessment, 22,* 287-303.

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

Shelton, N. R., Altwerger, B., & Jordan, N. (2009). Does DIBELS put reading first? *Literacy Research and Instruction, 48,* 137-148.

Shinn, M. (1995). Best practices in curriculum-based measurement and its use in a problem-solving model. In J. Grimes & A. Thomas (Eds.), *Best Practices in School Psychology III* (pp. 547-568). Silver Spring, MD: National Association of School Psychologists.

Shinn, M. R. (2007). Identifying students at risk, monitoring performance, and determining eligibility within response to intervention: Research on educational need and benefit from academic intervention. *School Psychology Review, 36,* 601-617.

Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21,* 459-479.

Skiba, R. S., Poloni-Staudinger, L., Gallini, S., Simmons, A. B., & Feggins-Aziz, R. (2006). Disparate access: The disproportionality of African American students with disabilities across educational environments. *Exceptional Children, 72,* 411-424.

Skiba, R. S., Simmons, A. B., Ritter, S., Gibb, A. C., Karega Rausch, M., Cuadrado, J., & Chung, C.-G. (2008). Achieving equity in special education: History, status, and current challenges. *Exceptional Children, 74,* 264-288.

South Carolina Education Accountability Act (1998) Section 59-18-300. Retrieved from http://www.scstatehouse.gov/code/t59c018.php

Tindal, G., Marston, D., & Deno, S. L. (1983). *The reliability of direct and repeated measurement* (Research Rep. No. 109). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly, 45,* 270-291.

VanDerHeyden,,, A. M., Witt, J. C., & Gilbertson, D. (2007). A multi-year evaluation of the effects of a Response to Intervention (RTI) model on identification of children for special education. *Journal of School Psychology, 45,* 225-256.

Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research & Practice, 18,* 137-146.

Watkins, M. W., & Hetrick, C. J. (1999). MacPotthoff: Automated calculation of the Potthoff regression bias procedure. *Behavior Research Methods, Instruments, & Computers, 31,* 710-711.

Weiss, L. G. & Prifitera, A. (1995). An evaluation of differential prediction of WIAT achievement scores from WISC-III FSIQ across ethnic and gender groups. *Journal of School Psychology, 33,* 297-304.

Witt, J. (2007). STEEP CBM screening and intervention for at risk children (Data file and code book). Retrieved from http://www.joewitt.org/.

APPENDIX A

TABLES

Table A1

Demographic Characteristics by Grade

	Grade			
	2	3	4	5
Total	240	217	215	162
Race				
Caucasian	165	149	148	131
African American	75	68	67	31
Gender				
Male	121	111	104	74
Female	119	106	111	88
Lunch Status				
Free Lunch	125	113	111	63
Reduced Lunch	28	21	19	9
Full-Pay Lunch	86	83	85	88
Education Status				
Regular Education	205	179	179	140
Special Education	35	38	36	22

Note. Special Education includes all categories of eligibility

Table A2

Means and Standard Deviations for Each Measure for Grade 2

		Fall ORF		Spring ORF		MAP	
Assessment	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Total	240	61.7	29.65	88.66	36.89	189.60	14.16
Race							
Caucasian	165	63.07	29.06	96.79	36.75	193.52	13.05
African American	75	58.72	30.90	70.76	30.51	180.96	12.65
Gender							
Male	121	60.14	29.97	85.77	36.30	189.48	15.31
Female	119	63.31	29.37	91.60	37.41	189.71	12.95
Lunch Status							
Free Lunch	125	61.49	29.09	76.84	34.06	185.14	13.84
Reduced Lunch	28	55.82	32.22	80.57	28.42	189.61	12.68
Full-Pay Lunch	86	64.50	29.38	108.72	35.19	196.03	12.73

Table A3

Means and Standard Deviations for Each Measure for Grade 3

		Fall ORF		Spring ORF		MAP	
Assessment	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Total	217	86.53	36.01	100.71	34.58	199.18	12.70
Race							
Caucasian	149	91.07	36.53	109.07	34.51	201.77	12.20
African American	68	76.60	32.96	82.39	26.96	193.49	11.99
Gender							
Male	111	86.04	35.26	97.98	35.06	198.35	13.79
Female	106	87.06	36.94	103.57	33.99	200.04	11.45
Lunch Status							
Free Lunch	113	80.56	33.91	92.05	30.47	196.17	12.22
Reduced Lunch	21	83.19	38.49	92.24	34.60	198.14	16.66
Full-Pay Lunch	83	95.52	36.73	114.64	35.65	203.53	11.03

Table A4

Means and Standard Deviations for Each Measure for Grade 4

		Fall ORF		Spring ORF		MAP	
Assessment	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Total	215	83.53	32.89	114.85	38.05	210.38	14.20
Race							
Caucasian	148	86.80	33.73	125.25	35.98	214.25	12.74
African American	67	76.30	29.93	91.88	32.13	201.82	13.56
Gender							
Male	104	79.50	29.00	109.19	36.81	209.00	14.71
Female	111	87.31	35.87	120.15	38.60	211.67	13.64
Lunch Status							
Free Lunch	111	78.23	26.68	103.52	36.31	205.57	14.39
Reduced Lunch	19	87.21	37.26	105.74	40.62	209.26	15.48
Full-Pay Lunch	85	89.64	38.05	131.68	33.67	216.91	10.82

Table A5

Means and Standard Deviations for Each Measure for Grade 5

		Fall ORF		Spring ORF		MAP	
Assessment	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Total	162	116.30	35.09	132.44	35.43	219.01	11.40
Race							
Caucasian	131	120.57	34.24	140.64	30.89	221.90	8.98
African American	31	98.23	33.31	97.77	32.69	206.77	12.49
Gender							
Male	74	115.96	32.41	133.47	35.81	219.04	11.33
Female	88	116.58	37.37	131.57	35.29	218.98	11.51
Lunch Status							
Free Lunch	63	111.06	34.89	119.46	35.22	214.59	12.05
Reduced Lunch	9	107.44	20.13	126.22	28.26	211.11	9.14
Full-Pay Lunch	88	121.35	36.06	141.49	33.40	222.86	9.52

Table A6

t Tests Comparing Gender, Race, and Lunch Status Group Means

	Fall DIBELS ORF		Spring DIBELS ORF		MAP	
Demographic Comparisons	*t*	*p*	*t*	*p*	*t*	*p*
			Grade 2			
Caucasian vs. African American	1.03	.305	5.74	.000**	7.06	.000**
Male vs. female	-.83	.409	-1.23	.222	-.13	.898
Free vs. reduced lunch	.86	.398	-.60	.5549	-1.66	.105
Reduced vs. full-pay lunch	-1.26	.213	-4.28	.000**	-2.33	.024*
Free vs. full-pay lunch	-.74	.463	-6.55	.000**	-5.89	.000**
			Grade 3			
Caucasian vs. African American	2.90	.004	6.17	.000**	4.70	.000**
Male vs. female	-.21	.835	-1.19	.235	-.98	.327
Free vs. reduced lunch	-.29	.772	-.02	.981	-.52	.609
Reduced vs. full-pay lunch	-1.32	.196	-2.63	.013*	-1.41	.172
Free vs. full-pay lunch	-2.91	.004*	-4.66	.000**	-4.41	.000**
			Grade 4			
Caucasian vs. African American	2.29	.024*	6.79	.000**	6.34	.000**
Male vs. female	-1.76	.080	-2.13	.034*	-1.38	.170
Free vs. reduced lunch	-1.01	.325	-.22	.826	-.971	.341
Reduced vs. full-pay lunch	-.26	.800	-2.59	.016*	-2.04	.053
Free vs. full-pay lunch	-2.36	.020*	-5.61	.000**	-6.30	.000**
			Grade 5			
Caucasian vs. African American	3.34	.002**	6.64	.000**	6.36	.000**
Male vs. female	-.11	.910	.34	.735	.04	.972
Free vs. reduced lunch	.45	.658	-.65	.528	1.02	.327
Reduced vs. full-pay lunch	-1.80	.094	-1.52	.159	-3.66	.005**
Free vs. full-pay lunch	-1.76	.080	-3.87	.000**	-4.53	.000**

** $p < .01$ level (2-tailed)
*$p < .05$ level (2-tailed)

Table A7

Intercorrelations Among All Measures for the Grade 2 Total Sample

	2.	3.
1. MAP	.08	.76**
2. Fall DIBELS ORF	1	.14*
3. Spring DIBELS ORF		1

** $p < .01$ level (2-tailed)
*$p < .05$ level (2-tailed)

Table A8

Intercorrelations Among All Measures for the Grade 3 Total Sample

	2.	3.
1. MAP	.08	.69**
2. Fall DIBELS ORF	1	.23**
3. Spring DIBELS ORF		1

** $p < .01$ level (2-tailed)
*$p < .05$ level (2-tailed)

Table A9

Intercorrelations Among All Measures for the Grade 4 Total Sample

	2.	3.
1. MAP	.21**	.69**
2. Fall DIBELS ORF	1	.29**
3. Spring DIBELS ORF		1

** $p < .01$ level (2-tailed)
*$p < .05$ level (2-tailed)

Table A10

Intercorrelations Among All Measures for the Grade 5 Total Sample

	2.	3.
1. MAP	.16*	.70**
2. Fall DIBELS ORF	1	.28**
3. Spring DIBELS ORF		1

** $p < .01$ level (2-tailed)
*$p < .05$ level (2-tailed)

Table A11

Pearson Product-Moment Correlation Coefficients Between DIBELS ORF and MAP Reading Comprehension Scores for Grade 2

	N	DIBELS ORF Fall	DIBELS ORF Spring
Total	240	.08	.76**
Race			
Caucasian	165	.12	.73**
African American	75	-.10	.72**
Gender			
Male	121	.06	.81**
Female	119	.10	.71**
Lunch Status			
Free Lunch	125	.03	.78**
Reduced Lunch	28	-.21	.74**
Full-Pay Lunch	86	.22*	.65**

** $p < .01$ level (2-tailed)
*$p < .05$ level (2-tailed)

Table A12

Pearson Product-Moment Correlation Coefficients Between DIBELS ORF and MAP Reading Comprehension Scores for Grade 3

	N	DIBELS ORF Fall	DIBELS ORF Spring
Total	217	.08	.69**
Race			
Caucasian	149	.05	.67**
African American	68	-.04	.62**
Gender			
Male	111	.02	.74**
Female	106	.15	.63**
Lunch Status			
Free Lunch	113	.14	.69**
Reduced Lunch	21	-.19	.74**
Full-Pay Lunch	83	-.06	.63**

** $p < .01$ level (2-tailed)
* $p < .05$ level (2-tailed)

Table A13

Pearson Product-Moment Correlation Coefficients Between DIBELS ORF and MAP Reading Comprehension Scores for Grade 4

	N	DIBELS ORF Fall	DIBELS ORF Spring
Total	215	.21**	.69**
Race			
Caucasian	148	.22**	.60**
African American	67	.04	.68**
Gender			
Male	104	.28**	.71**
Female	111	.14	.66**
Lunch Status			
Free Lunch	111	.14	.70**
Reduced Lunch	19	-.09	.57*
Full-Pay Lunch	85	.29**	.56**

** $p < .01$ level (2-tailed)
*$p < .05$ level (2-tailed)

Table A14

Pearson Product-Moment Correlation Coefficients Between DIBELS ORF and MAP Reading Comprehension Scores for Grade 5

	N	DIBELS ORF Fall	DIBELS ORF Spring
Total	162	.16*	.70**
Race			
Caucasian	131	.06	.58**
African American	31	-.04	.71**
Gender			
Male	74	.07	.66**
Female	88	.23*	.74**
Lunch Status			
Free Lunch	63	.09	.70**
Reduced Lunch	9	.43	.81**
Full-Pay Lunch	88	.14	.65**

** $p < .01$ level (2-tailed)
* $p < .05$ level (2-tailed)

Table A15

Results for Stepwise Multiple Regression Analyses Predicting MAP Reading Comprehension for Grade 2

Model	*R*	R^2	*Adj.* R^2	Change statistics		
				ΔR^2	ΔF	Sig. *F* change
DIBELS ORF Spring	.76	.58	.58	.58	327.77	.000
DIBELS ORF + Race	.78	.61	.61	.03	17.95	.000

Table A16

Coefficients for Significant Predictor Variables for Multiple Regression Analysis Predicting MAP Reading Comprehension for Grade 2

Variable	*B*	*SEB*	*β*	*t*	*P*
DIBELS ORF spring	.27	.02	.70	16.31	.000
Race	-5.55	1.31	-.18	-4.24	.000

Table A17

Results for Stepwise Multiple Regression Analyses Predicting MAP Reading Comprehension for Grade 3

Model	*R*	R^2	*Adj.* R^2	Change statistics		
				ΔR^2	ΔF	Sig. *F* change
DIBELS ORF Spring	.69	.48	.48	.48	198.10	.000

Table A18

Coefficients for Significant Predictor Variables for Multiple Regression Analysis Predicting MAP Reading Comprehension for Grade 3

Variable	*B*	*SEB*	*β*	*t*	*P*
DIBELS ORF spring	.25	.02	.69	14.08	.000

Table A19

Results for Stepwise Multiple Regression Analyses Predicting MAP Reading Comprehension for Grade 4

Model	*R*	R^2	*Adj.* R^2	Change statistics		
				ΔR^2	ΔF	Sig. *F* change
DIBELS ORF Spring	.69	.47	.47	.47	190.59	.000
DIBELS ORF + Race	.70	.49	.49	.02	8.92	.003

Table A20

Coefficients for Significant Predictor Variables for Multiple Regression Analysis Predicting MAP Reading Comprehension for Grade 4

Variable	*B*	*SEB*	*β*	*t*	*P*
DIBELS ORF spring	.24	.02	.64	12.45	.000
Race	-4.35	1.46	-.15	-2.99	.003

Table A21

Results for Stepwise Multiple Regression Analyses Predicting MAP Reading Comprehension for Grade 5

Model	*R*	R^2	*Adj.* R^2	Change statistics		
				ΔR^2	ΔF	Sig. *F* change
Spring ORF	.70	.50	.49	.50	157.32	.000
Spring ORF + Race	.74	.54	.54	.05	15.82	.000
Spring ORF + Race + Lunch Status	.75	.56	.55	.02	6.95	.009

Table A22

Coefficients for Significant Predictor Variables for Multiple Regression Analysis Predicting MAP Reading Comprehension for Grade 5

Variable	*B*	*SEB*	*β*	*T*	*p*
Spring ORF	.19	.02	.58	9.69	.000
Spring ORF + Race	-7.03	1.73	-.24	-4.06	.000
Spring ORF + Race + Lunch Status	-6.90	2.62	-.14	-2.64	.009

Table A23

F, df, and p values for Simultaneous Slope and Intercept Comparisons Between Demographic Groups in Predicting MAP Reading Comprehension scores Using Fall DIBELS ORF and Spring DIBELS ORF for Grade 2

	Fall DIBELS ORF			Spring DIBELS ORF		
Demographic Comparisons	*F*	*df*	*p*	*F*	*df*	*p*
Caucasian vs. African American						
Simultaneous test	25.29**	2, 236	.000	9.39**	2, 236	.000
Slope test	2.65	1, 236	.105	.84	1, 236	.359
Intercept test	2.73	1, 236	.100	6.48*	1, 236	.012
Males vs. females						
Simultaneous test	.04	2, 236	.962	5.39**	2, 236	.005
Slope test				9.18**	1, 236	.003
Intercept test				5.35*	1, 236	.022
Free lunch vs. reduced lunch vs. full-pay lunch						
Simultaneous test	9.48	4, 234	.138	2.51	4, 234	.053
Slope test						
Intercept test						

** $p < .01$ level (2-tailed)
* $p < .05$ level (2-tailed)

Table A24

F, df, and p values for Simultaneous Slope and Intercept Comparisons Between Demographic Groups in Predicting MAP Reading Comprehension scores Using Fall DIBELS ORF and Spring DIBELS ORF for Grade 3

	Fall DIBELS ORF			Spring DIBELS ORF		
Demographic Comparisons	*F*	*df*	*p*	*F*	*df*	*p*
Caucasian vs. African American						
Simultaneous test	10.29**	2, 213	.000	1.07	2, 213	.344
Slope test	.28	1, 213	.595			
Intercept test	1.64	1, 213	.202			
Males vs. females						
Simultaneous test	.81	2, 213	.447	2.45	2, 213	.089
Slope test						
Intercept test						
Free lunch vs. reduced lunch vs. full-pay lunch						
Simultaneous test	4.97	4, 211	.159	2.41*	4, 211	.023
Slope test				3.82	2, 211	.380
Intercept test				3.96**	2, 211	.000

** $p < .01$ level (2-tailed)

* $p < .05$ level (2-tailed)

Table A25

F, df, and p values for Simultaneous Slope and Intercept Comparisons Between Demographic Groups in Predicting MAP Reading Comprehension scores Using Fall DIBELS ORF and Spring DIBELS ORF for Grade 4

	Fall DIBELS ORF			Spring DIBELS ORF		
Demographic Comparisons	*F*	*df*	*p*	*F*	*df*	*p*
Caucasian vs. African American						
Simultaneous test	19.37**	2, 211	.000	5.36**	2, 211	.005
Slope test	1.12	1, 211	.291	2.65	1, 211	.105
Intercept test	1.59	1, 211	.208	6.22*	1, 211	.013
Males vs. females						
Simultaneous test	1.67	2, 211	.190	.82	2, 211	.440
Slope test						
Intercept test						
Free lunch vs. reduced lunch vs. full-pay lunch						
Simultaneous test	8.22	4, 209	.396	3.79	4, 209	.058
Slope test						
Intercept test						

** $p < .01$ level (2-tailed)
* $p < .05$ level (2-tailed)

Table A26

F, df, and p values for Simultaneous Slope and Intercept Comparisons Between Demographic Groups in Predicting MAP Reading Comprehension scores Using Fall DIBELS ORF and Spring DIBELS ORF for Grade 5

	Fall DIBELS ORF			Spring DIBELS ORF		
Demographic Comparisons	*F*	*df*	*p*	*F*	*df*	*p*
Caucasian vs. African American						
Simultaneous test	27.33**	2, 158	.000	10.38**	2, 158	.000
Slope test	.27	1, 158	.606	4.59*	1, 158	.034
Intercept test	3.40	1, 158	.067	11.04**	1, 158	.001
Males vs. females						
Simultaneous test	.36	2, 158	.701	.44	2, 158	.644
Slope test						
Intercept test						
Free lunch vs. reduced lunch vs. full-pay lunch						
Simultaneous test	6.53	4, 156	.686	3.95	4, 156	.304
Slope test						
Intercept test						

** $p < .01$ level (2-tailed)
* $p < .05$ level (2-tailed)