8-15-2013

Test Nationally, Benchmark Locally: Using Local DIBELS Benchmarks to Predict Performance on the PSSA

Matthew R. Ferchalk
Indiana University of Pennsylvania

TEST NATIONALLY, BENCHMARK LOCALLY:

USING LOCAL DIBELS BENCHMARKS TO PREDICT PERFORMANCE

ON THE PSSA

A Dissertation

Submitted to the School of Graduate Studies and Research

in Partial Fulfillment of the

Requirements for the Degree

Doctor of Education

Matthew R. Ferchalk

Indiana University of Pennsylvania

August 2013

Indiana University of Pennsylvania
School of Graduate Studies and Research
Department of Educational and School Psychology

We hereby approve the dissertation of

Matthew R. Ferchalk

Candidate for the degree of Doctor of Education

June 24, 2013	Signature on File
	Timothy J. Runge, Ph.D. Assistant Professor of Educational and School Psychology, Advisor
June 24, 2013	Signature on File
	Joseph F. Kovaleski, D.Ed. Professor of Educational and School Psychology
June 24, 2013	Signature on File
	Mark J. Staszkiewicz, Ed.D. Professor of Educational and School Psychology
June 24, 2013	Signature on File
	David Lillenstein, D.Ed. School Psychologist Derry Township School District

ACCEPTED

Signature on File	
Timothy P. Mack, Ph.D. Dean School of Graduate Studies and Research	

Title: Test Nationally, Benchmark Locally: Using Local DIBELS Benchmarks to Predict Performance on the PSSA

Author: Matthew R. Ferchalk

Dissertation Chair: Timothy J. Runge, Ph.D.

Dissertation Committee Members: Joseph F. Kovaleski, D.Ed.
Mark J. Staszkiewicz, Ed.D.
David J. Lillenstein, D.Ed.

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) benchmarks are frequently used to make important decision regarding student performance. More information, however, is needed to understand if the nationally-derived benchmarks created by the DIBELS system provide the most accurate criterion for evaluating reading proficiency. The DIBELS benchmarks are calculated based on performance with a nationally-normed standardized achievement test (Good et al. 2011b). Therefore, it may not accurately represent the standard of performance on a state assessment. The DIBELS benchmarks may show a high number of false positives or false negatives when the benchmark is used to predict state test scores. A different criterion that more accurately reflects local expectations may be needed. A locally-generated benchmark expectation, established using a state assessment as a criterion for success, may provide a valid alternative to the DIBELS benchmark.

DIBELS Oral Reading Fluency (ORF) data and Pennsylvania System of School Assessment (PSSA; PDE, 2010) scores were collected from two school districts in Pennsylvania. The collected data reflected fall, winter, and spring DIBELS ORF scores for students in grades 3 through 5 as well as their scores on the PSSA. Using logistic regression, locally-generated ORF benchmarks using PSSA performance as the criterion for successful outcomes were created for both school districts. Diagnostic accuracy statistics, including sensitivity, specificity, negative predictive power, and positive predictive power, and overall accuracy percentage as well as values for kappa and phi were calculated for each set of benchmarks. Contrary to the hypothesis, significant differences were not found between the locally-generated benchmarks and the DIBELS benchmarks in PSSA prediction accuracy. Significant differences between the locally-generated benchmarks and the DIBELS benchmarks levels of sensitivity, specificity, and positive predict power was produced in both school districts. Given these differences, the use of more than one set of benchmarks scores for instructional decision making is recommended. The author also recommends that school psychologists should learn how well the nationally-derived DIBELS benchmark corresponds with the local

expectations to ensure that sound decision practices are used when determining how to best meet student needs.

ACKNOWLEDGMENTS

I would like to thank everyone who helped make this dissertation possible. I would first like to thank my wonderful wife Dr. Joie Cogan-Ferchalk. You are my source of love, support, inspiration, motivation, and even a little competition (you win by 1 month). I espcially would like to thank you for your incredible patience as you endured my long hours of “slight” iritibility as I sat behind the computer. You are so amazing and I love you very much!

I would also like to thank my dissertation chair Dr. Timothy Runge. You gave me the extra push that I needed to make it to the top of this mountain. I will be forever greatful! To my excellent committee members, Dr. David Lillenstein, Dr. Joseph Kovaleski, and Dr. Mark Staszkiewicz: Thank you so much for your guidance, recommendations, and suggestions. You have all helped to make this dissertation the best it could be and I greatly appreciate it!

Special thanks goes to my parents Joseph and Sheila Ferchalk. Thank you for the love and support you have given me over my lifetime and for the sacrifices that you made so that I could succeed. Without your guidance, this dissertation would never have made it past page 1. This doctorate belongs to you as much as it does me.

This dissertation would not have been possible without my and collegues who have helped me, in their own way, to achieve this goal. Dr. Richard Hall: Thank you for your guidance throughout this process. Your support has helped me make it through to the finish line. Dr. Jason Pedersen: Thanks for giving me an extra boost of motivation when I needed it the most. Though I lost the bet, I know that my Franco Harris jersey has gone to a good home. My fellow Northern Lebanon school psychologist, Fiona Richardson: Thank you for your help with data collection and for your support. I hope to do the same for you someday. Mike Dunkel: I owe you for our Wednesday evening "therapy" sessions. Our decompression time helped me make it through the next day of writing. Jim Walter: Thank you for your friendship, support, and for making sure that my ego never gets too big. And lastly, Caitlin Flinn: Thanks for serving as my counselor and my cheering section. It will be your turn soon!

Working in education over the past 7 years has made me realize how the good work of teachers often goes unappreciated. I would therefore like to show my appreciation for my 9th and 11th grade algebra teacher, Mrs. Myra Whysong-Krentz. As I worked through hours data analysis I realized that, without your guidance 20 years ago, Chapter 4 would not have been possible.

To my baby boy, Jonah: Your upcoming arrival has served as the greatest possible "dissertation completion" motivation I could have asked for. Now that these 5 chapters are written, I will turn the page and start a new and more exciting chapter and I couldn't be happier! I can't wait to meet you!

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

In a continued push toward accountability in education as a result of the regulations derived from No Child Left Behind (NCLB; 2001), statewide testing results have increasing consequences for students, teachers, and school districts (Deno, 2003; Shapiro, Solari, & Petscher 2008). In some states, student scores on accountability-driven state assessments have been linked to teacher evaluations, salary increases, and in some cases have been cause for teacher termination (Amrein & Berliner, 2002). Sanctions, such as loss of autonomy in decision making, may be enforced by the state department of education level to schools who fall below the accepted standards (Braden & Tayrose, 2008). In other states, monetary bonuses are given to teachers and administrators who show the greatest student gains. Even in rare cases, rewards and punishments have been handed out to individual students for their successes and failures on the state test (Amrein & Berliner, 2002; Braden & Tayrose, 2008).

Given the importance of these high-stakes assessments, educators have sought ways of helping students become successful readers and have their efforts reflected on the state-wide test (Silberglitt & Hintze, 2005). A particular emphasis on the implementation of research-based strategies to

improve reading has received considerable attention in both the legal statutes and the research literature (McGlinchey & Goodman, 2008; Reyna, 2004). Noted by Tilly (2008) the phrase *scientific research-based practice* appears in the text of NCLB (2001) on 111 separate occasions. The Individuals with Disabilities Education Improvement Act (IDEIA; 2004) also prioritizes the use of research-based practices to assist in the identification of students with learning disabilities. In addition, The National Reading Panel (NRP) was established in 1997 to assess the effectiveness of a wide variety of methods for teaching children to read (NICHD, 2000). The research findings reported by NRP have provided educators with a valid frame of reference for the application of empirical research to sound instructional practices. Through a comprehensive meta-analysis, the NRP identified five areas of reading that are the most effective when targeted in beginning reading instruction (Ehri, 2004). These "Five Big Ideas" include phonemic awareness, knowledge of the alphabetic principal (phonics), reading fluency, vocabulary development, and strategies to increase reading comprehension. Direct instruction efforts, within the context of these Five Big Ideas, have been consistently linked to increased reading achievement (NICHD, 2000).

Research-based strategies can take many different forms and may be implemented across various settings. They have employed numerous instructional practices that are derived from differing philosophies on reading development including whole language approaches, direct phonics instruction, and a balance of the two paradigms (Casey & Howe, 2002). Within these frameworks, teachers have engaged in a wide variety of instructional practices including partner reading and other peer assisted strategies, sight word vocabulary practice, guided reading groups, differentiated instruction, cooperative learning, learning style approaches and many other educational innovations (Ellis, 2001). These supports have been provided by the classroom teacher in homogeneous skill groups, small group instruction with the building reading specialist or classroom aide, parent volunteers, and peer tutors.

Many school districts have employed a Response to Intervention (RtI) model to guide decisions about the provision of these instructional supports (Glover & DiPerna, 2007). This model pulls together many of these research-based interventions under one comprehensive umbrella that covers a wide variety of applications but is most commonly used for structuring reading, math, and behavior interventions (Glover & DiPerna, 2007; Kovaleski, 2007). RtI is comprised of a few basic components including multi-tier service delivery, the

provision of evidence-based interventions, formative assessment, and data-based decision making (Glover & DiPerna, 2007). Through these components, students are provided evidence-based interventions that are designed to deliver academic skills at a level of intensity targeted to meet individual student needs. The determination of which interventions are used and the level of intensity needed are based on the results and analysis of ongoing formative assessments (Daley, Martens, Barnett, Witt, & Olsen, 2007; Lichtenstein, 2008; Tilly, 2006).

In addition, formative assessments are also used to adjust the interventions and instruction to ensure positive outcomes for students (Glover & DiPerna, 2007; Kovaleski, 2007). Although positive outcomes will likely be measured by state test scores, these once per year assessments do not provide sufficient information for educators to ensure that the necessary steps are taken improve student performance (Amrein & Berliner, 2002; Medina & Riconscente, 2006). For each subject area only a limited number of questions are presented to the student. The negligible amount of data produced by the test gives educators only minimal information that can be used to adjust instruction. The timing of the state proficiency test creates a second important problem. Students are not assessed until the end of each school year

ensuring their results will not be released until the summer holiday. By the time teachers receive the outcomes of the tests it is far too late to address the needs of the students who fell below proficiency. When the new school year begins, last year's students will have moved on to a new teacher and their old desks will have been filled by an entirely new group of children.

It is ironic that the most crucial test taken by students during the school year provides only a minimal amount of information that can be used to improve instructional practices. It is therefore essential for educators to assess student performance throughout the school year and use the information to adjust their instruction. When accomplished, a student's chance of meeting proficiency will greatly improve (Shapiro, 2008). To meet this goal, many school districts have chosen to universally screen all students using curriculum-based measurement (CBM) throughout the school year (Crawford, Tindal, & Stieber, 2001; Deno, 1985; Hintze & Silberglitt, 2005; Shapiro, Solari, & Petscher 2008; Silberglitt & Hintze, 2005). Conducting CBM can provide educators with useful information that can be used to make necessary changes to instruction. By adjusting or fine-tuning the curriculum and instruction throughout the academic year, school personnel can help students become successful readers

who will be better prepared to demonstrate their skills on the state assessment (Silberglitt & Hintze, 2005).

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good et al., 2011a), AIMSweb (Shinn & Garmin, 2006), and Easy CBM (Alonzo, Tindal, Ulmer, & Glasgow, 2006) are sets of commercially produced curriculum-based measures that have been adopted by many school districts across the country for the purpose of universal screening in reading skill acquisition. They are designed to efficiently screen a student population, identify those students in need of reading support, and monitor the progress of struggling readers toward successful reading outcomes. Each of these assessment packages measure student skills across a variety of subtests including passage reading followed by multiple choice questions, oral reading fluency (ORF), cloze reading procedures, and retell fluency. Of these assessments, ORF has been an essential component of CBM since its inception and a great deal of research has been dedicated to its correlation with overall reading performance (Deno, 1985; Hintze & Silberglitt, 2005).

ORF is designed to efficiently assess a student's reading fluency which is defined as the effortless, automatic ability to read words in connected text (Stahl, 2004). Fluency is important to emergent reading skills because "if children can

recognize words quickly and automatically, then word recognition does not interfere with comprehension, and children can understand any text within their language ability" (Stahl, 2004, p. 189). This connection between reading fluency and reading comprehension is supported by decades of research. Numerous studies have shown strong correlations between ORF and reading comprehension on a wide variety of assessments (Baker et al., 2008; Deno, Mirkin, & Chiang, 1982; Fuchs, Fuchs, & Maxwell, 1998; Hintze & Silberglitt, 2005; Hosp & Fuchs, 2005; Jenkins & Jewell, 1993; Shinn, Good, Knutson, Tilly, & Collins, 1992). Given this strong connection, measures of ORF appear to be a highly valid method to use when screening students for difficulties in reading comprehension.

Universal screening with DIBELS ORF is typically assessed three times per year in the fall, winter and spring (Good et al., 2011b). During each assessment, students are typically asked to read three short passages for 1 minute. Of the three scores, the median number of words that are read correctly in 1 minute (wcpm) is selected as the score that will represent the student's level of oral reading fluency. The ORF scores earned by students on the DIBELS are then compared to benchmark expectations, or cut scores, which indicate the likelihood that a student will reach subsequent reading goals.

Students who score at or above the benchmark are identified as *Likely to Need Core Support* or *Low-Risk*. According to the DIBELS system, the odds that a student who achieves a score in this area will achieve early literacy goals are 80 - 90% (Good et al., 2011b). Students who score below the benchmark goal fall in one of two categories. First, students that fall below the benchmark expectation but above the cut point for risk are identified as *Some-Risk* or *Likely to Need Strategic Support*. According to the authors, the odds that a student who falls in this category will reach early reading goals are between 40 - 60%. Students who fall well below the cut point for risk are identified as *At-Risk* or *Likely to Need Intensive Support*. The odds that a student labeled *At-Risk* will reach early literacy goals are between 10 - 20%.

Teachers and administrators frequently utilize these benchmark expectations to make important decisions regarding children's programming and needs (Shapiro, 2008). Through the application of the benchmarks, educators can decide to re-teach skills and concepts, provide small group instruction, focus on basic skill attainment, increase or reduce intervention time, or determine if additional interventions are needed to improve student achievement (Ikeda, Neessen, & Witt, 2008; VanDerHayden, 2011). As students progress throughout the school year, their response to the instruction

in relation to the benchmark expectations may be used as an important indicator to determine whether or not special education services are needed (Dynamic Measurement Group, 2010; Shapiro, 2008).

An important goal of the application of CBM is to use the data to employ interventions that will enable students to become proficient readers on the end of the year assessment. Naturally, this has lead schools to use CBM data to predict performance on the state assessment (Silberglitt, 2008). Within the last decade, numerous studies have been published which have measured the connection between CBM and state assessments (Crawford, Tindal, & Steiber , 2001; Good, Simmons & Kame'enui, 2001; McGlinchey & Hixon, 2004; Merino & Beckman, 2010; State & Jacobson, 2002). The results routinely find strong relationships between CBM and state assessments regardless of which state assessment was given.

Unfortunately, complications can arise when applying CBM as a tool to predict state testing performance. Although a strong correlation exists between ORF and performance on state assessments, it is not a perfect correlation. Because of this imperfection, errors in the predictions of state test proficiency from CBM are inevitable. This can be particularly problematic where benchmark scores are concerned. The number of false positives (incorrect prediction of a reading deficit)

and false negatives (incorrect prediction of a proficient student) will vary depending on where the benchmark score cut scores are drawn (Silberglitt & Hintze, 2005). An increase in the benchmark score will decrease the number of false negatives but will simultaneously increase the number of false positives (Silberglitt, 2008). The reverse is true when benchmark scores are lowered. This error in prediction can complicate the high-stakes decisions that determine what services and supports students receive. It is possible that students could be denied reading supports after earning a proficient benchmark score. It is also possible that students could receive unnecessary time intensive and costly interventions because they earned an inaccurate below benchmark ORF score.

The authors of DIBELS consider the problem of prediction accuracy when developing their benchmarks. They employ a step-by-step process designed to carefully balance the number of false positives and false negatives (Good et al., 2011b). After this process is completed, the benchmarks should reflect adequate reading outcomes that can be measured on a variety of assessments; however, in their admirable desire to create a tool which can be utilized on a national scale, the developers of DIBELS may have sacrificed diagnostic accuracy at the individual state, district, or elementary school building

level (Silberglitt, 2008). This is because the local expectations are not necessarily reflected by a nationally normed benchmark score. According to Kingsbury, Olso, Cronin, Hauser, and Houser (2004), there can be a great disparity between the level of difficulty from one state proficiency test to the next. They found large discrepancies across states between the score on a nationally normed test of achievement needed to accurately predict proficiency on a state assessment. For example, the minimum proficiency in South Carolina fell at the 67th percentile and at the 51st percentile in California. In Minnesota, Oregon, Idaho, Montana, Indiana, and Illinois, proficiency fell at the 35th percentile. Students in Colorado and Texas students can meet proficiency with scores equivalent to the 13th percentile. These discrepancies exist because each state has uniquely addressed the regulations and requirements under NCLB. In the absence of a national assessment or curriculum, each state has developed unique requirements for proficiency and individualized tests to measure satisfactory reading outcomes. This creates a situation where the level of difficulty of the assessment and/or the expectation for proficiency may be higher in one state than it is in another.

Using a nationally normed assessment like DIBELS to monitor progress toward the end of the year proficiency

assessment is complicated by these inter-state differences. Consider this scenario: The third-grade benchmark expectation for *Low-Risk* reading performance is 70 wcpm. This suggests that a student who is reading above this benchmark score possesses adequate general reading skills, which will be reflected on the state assessment; however, the DIBELS benchmarks are calculated based on performance with a nationally normed standardized achievement test and may not accurately represent the standard of performance on a state assessment. The result may show a high number of false positives and false negatives when using the benchmark score from one state to the next. This could mean that students who attend school in a state with a higher standard or more difficult test, such as California or South Carolina, may not meet proficiency even after earning a DIBELS score that suggested *Low-Risk* performance. Conversely, a third-grade student who earns a below benchmark ORF score may still meet proficiency in Texas and Colorado where the standard for proficiency is much lower.

In the absence of a national curriculum or national assessment, measures like DIBELS and other universal screeners may open a window into what proficient reading looks like across the country. They are based on a nationally normed reading assessment and are designed to be accurate predictors

of general reading outcomes independent of curricula and individual state standards (Good et al., 2011b). Because of the discrepancies from one state to the next, school districts should avoid simply employing a nationally normed set of benchmarks that may not be connected to their local expectations (Silberglitt, 2009). If a school district desires to know how their students are performing in relation to their own state expectations, then a different locally-generated metric may be needed.

Statement of the Problem

CBM in oral reading fluency share a high correlation with general reading outcomes (Baker et al., 2008; Deno, Mirkin, & Chiang, 1982; Hintze & Silberglitt, 2005; Hosp & Fuchs, 2005; Jenkins & Jewell, 1993; Shinn et al., 1992). Because of this strong relationship, ORF has been used as a reliable tool to identify students who are at-risk for reading failure as well as monitor their progress toward successful reading outcomes. Further, it can be used as a tool to help identify students with learning disabilities. The DIBELS system has created their benchmarks as a way to aid educators in making these decisions (Good, Kaminski, Dewey et al., 2011). By creating their benchmarks, they provided a simple and efficient way to compare student performance to well-researched and valid expectations for successful reading outcomes.

There is sufficient evidence, however, to suggest that the DIBELS benchmarks are not always consistent with local expectations (Kingsbury et al., 2004). Because of the error inherent within the benchmarks as well as the discrepancies in the expectations for proficiency from one state to another, the accuracy of the predictions made from the benchmark expectation to a state test may be questionable. Situations can occur where students earn scores above the DIBELS benchmarks yet still fall below proficiency on the state assessment. Conversely, students who fall below the DIBELS benchmarks may otherwise demonstrate successful reading skills on a variety of other indicators of reading performance, including a state proficiency assessment.

Given the gravity of the decisions made with the DIBELS benchmarks, more information is needed to understand if the nationally-derived benchmarks created by the DIBELS system are providing the most accurate criterion for reading proficiency or whether locally-generated benchmarks are more accurate. By analyzing the relationship with the PSSA, this dissertation will help to determine whether the DIBELS benchmarks are sufficient for predicting performance on a state proficiency assessment, or whether benchmarks generated at the local level should be considered. A discussion of the most appropriate

usage of both the locally-generated benchmark as well as the nationally normed benchmark will be provided.

Research Questions and Hypotheses

Research Question 1

Are there statistically significant mean differences on DIBELS benchmarks between the two participating schools? It is hypothesized that differences will not exist between the benchmark scores generated for the participating school districts. This hypothesis was made because the study sites share relatively similar racial, sex, and socio-economic demographic characteristics. In addition, analogous percentages of special education students and English Language Learners are present in both study sites.

Research Question 2

What are the correlations between the fall, winter, and spring DIBELS ORF scores and performance on the PSSA in grades 3 - 5? Separate correlations will be calculated for each grade and each assessment period (fall, winter, and spring).

The analysis of this research question is dependent upon the results of Research Question 1. Benchmark scores from the schools will be generated and correlated with PSSA in grades 3 – 5. If the two participating schools' benchmark scores are significantly different, as determined from the analyses associated with Research Question 1, then separate benchmarks

will be generated for both study sites. If significant differences are not identified, the data for both study sites will be combined to generate one set of scores.

Consistent with previous research (Baker et al., 2008; Deno, Mirkin, & Chiang, 1982; Hintz & Silberglitt, 2005; Hosp & Fuchs, 2005; Jenkins & Jewell, 1993; Shinn et al., 1992), it is hypothesized that ORF at each grade level will demonstrate moderate to strong correlations with the results of the PSSA.

Research Question 3

What are the locally-generated benchmark scores in the fall, winter, and spring in grades 3 - 5? Logistic regression will be used to calculate the locally-generated benchmarks. This procedure was chosen because the cut score in a range of student ORF scores that produces the highest percentage of correct predictions of the PSSA is selected and utilized as part of the analysis with logistic regression.

The analysis of this Research Question is dependent upon the results of Research Question 1. If significant differences are identified between the benchmark scores of participating sites, then separate benchmarks will be generated for both study sites. If significant differences are not identified, the data for both study sites will be combined to generate one set of benchmark scores.

It is hypothesized that the locally-generated benchmarks will be lower than those created by the DIBELS system. This is because the DIBELS benchmarks are designed to carefully balance the number of false positives and false negatives in an attempt to ensure that more students are identified as in need of additional supports (Good et al., 2011b). This inflated cut score, useful for screening purposes, sacrifices the accuracy of the prediction of PSSA proficiency. In addition, the logistic regression procedure used in this study maximizes the percentage of true positives only and produces a benchmark score that is not artificially inflated but maximizes the prediction accuracy on the PSSA. According to Hintze and Silberglitt (2005), logistic regression typically produces cut scores that are lower than other methods of calculation.

Research Question 4

Are the locally-generated benchmarks able to predict PSSA proficiency with significantly greater accuracy than the DIBELS benchmarks? Additionally, are measures of diagnostic accuracy (sensitivity, specificity, PPP, and NPP) significantly different based on the derivation of the benchmarks?

The analysis of this research question is dependent upon the results of Research Question 1. If significant

differences are identified between the benchmark scores, then benchmarks scores generated for both study sites will be compared separately with the DIBELS benchmark scores. If significant differences are not identified, then the benchmark scores for both study sites will be combined to generate one set of benchmark scores that will be compared to the DIBELS benchmarks.

It is hypothesized that significant differences will be identified between the locally-developed benchmarks and the DIBELS-generated benchmarks in their ability to reliably predict PSSA performance. It is further hypothesized that the locally-generated benchmarks will more accurately predict PSSA performance. In addition, significant differences will be present between the diagnostic accuracy statistics for both sets of benchmarks. These hypotheses are suggested for two reasons. First, the DIBELS benchmarks are designed to carefully balance the number of false positives and false negatives to create an inflated cut score that ensures more students are identified as in need of additional supports (Good et al., 2011b). This inflated score, however, will likely produce a less accurate prediction. Second, local benchmarks developed by Ferchalk, Richardson, and Cogan-Ferchalk (2010) more accurately predicted proficiency on the

PSSA than DIBELS generated benchmarks. Similar findings are predicted for this study.

Definition of Terms

Benchmark

Benchmarks are criterion referenced, research-supported goal scores which reflect satisfactory skill progress (Good, Kaminski, Dewey et al., 2011; Shapiro, 2008). They represent a standard for gauging student skill development (Kaminski, Cummings, Powell-Smith, & Good, 2008). Benchmarks may be referred to by other names including target scores, cut scores and benchmark expectations.

Curriculum-Based Measurement

Curriculum-based measurement (CBM) is a collection of typically fluency-based, assessments used to assess student progress in one of several academic skill areas (Deno, 1985; Shinn, 2008). They are standardized measures that can be assessed in a short amount of time, usually 1 - 5 minutes (Shinn, 2008). CBM provides reliable data that can be utilized for instructional decision making and program evaluation.

Diagnostic Accuracy Statistics

Interpretations of diagnostic accuracy statistics are derived from a medial model approach (Silberglitt, 2008; VanDerHayden, 2011). In this sense, a positive result will

indicate the presence of a reading problem as positive test in the medical field reflects the presence of a disease (Silberglitt, 2008; VanDerHayden, 2011). In both examples the word positive reflects an unfavorable outcome (Silberglitt, 2008). Similarly, the word negative reflects favorable results such as the lack of a reading problem or disease. A false positive indicates the number of students who are predicted to be at risk for reading failure who are not at risk on an outcome measure. Conversely, a true positive indicates the number of at-risk students who were accurately predicted to be at risk. A false negative describes the number of students who are predicted to be proficient but fail to meet proficiency on the outcome measure. A true negative refers to the number of **s**tudents who are predicted to be proficient and who meet proficiency on the outcome measure.

In addition to overall accuracy of prediction percentage, the following four diagnostic accuracy statistics are used to determine how precisely an ORF score will predict performance on a state test (Hintze & Silberglitt, 2005; Silberglitt & Hintze, 2005). The four statistics include sensitivity, specificity, positive predictive power, and negative predictive power. The number of false positives and true positives and false negatives and true negatives are uniquely balanced by each of these metrics.

Sensitivity. Sensitivity refers to the percentage of students who were correctly predicted to fail the outcome measure out of all of the students who actually failed the outcome (Silberglitt, 2008). It specifically refers to the proportion of true positives (VanDerHayden, 2011). A benchmark score with a high level of sensitivity will better identify students with reading concerns.

Specificity. Specificity refers to the percentage of students who were accurately predicted to pass the outcome out of all of the students who passed the outcome measure. It refers to the proportion of true negatives (VanDerHayden, 2011). A benchmark score with a high degree of sensitivity will more accurately identify students who do not show reading difficulties.

Positive predictive power. Positive predictive power (PPP) is the proportion of true positives or the percentage of students predicted to be positive that are actually positive on the outcome measure (VanDerHayden, 2011). It indicates the proportion of students that were predicted to fail and actually failed the state test (Silberglitt, 2008). PPP will be reported as a percentage.

Negative predictive power. Negative predictive power (NPP) indicates the proportion true negatives or the percentage of students predicted to be negative on the outcome

measure who are actually negative (VanDerHayden, 2011). It refers to the proportion of students that were predicted to pass who actually passed the state test (Silberglitt, 2008). NPP is also reported as a percentage.

Dynamic Indicators of Basic Early Literacy Skills

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) are a collection of brief fluency-based assessments used to assess early reading skill acquisition in grades 3 - 6 (Good et al., 2011a). They are designed to identify students who are having reading difficulties so that supports can be provided early to prevent later reading problems (Kaminski et al., 2008). In addition, DIBELS provides valuable information on the effectiveness of intervention efforts.

DIBELS composite score. The DIBELS composite score is an amalgamation of several separate indicators and, according to the authors, provides a more reliable measure of students' reading skills than any of the individual indicators (Good, Kaminski, Dewey et al., 2011). At the third through fifth grade levels, the DIBELS composite score is comprised of ORF including the number of words read correctly per minute (WCPM), the number of words that a student correctly retells after reading the passage (Retell Fluency), and the student's percentage of words read correctly (Accuracy Percentage). In

addition, student's score on the DAZE are added into the composite.

DIBELS Daze. Daze is a maze procedure that has been standardized by the DIBELS system to assess reading comprehension (Good et al., 2011). In a maze procedure, a predetermined number of words are deleted from a paragraph and are replaced with a multiple choice selection (McKenna & Stahl, 2003). Students are asked to choose the response which best within the context of the assessment. Maze assessments are designed to assess the reasoning skills that accompany reading comprehension and a student's ability to form meaning from what they have read (Shapiro, Solari, & Petscher, 2008).

DIBELS Oral Reading Fluency. DIBELS Oral Reading Fluency (ORF) is a curriculum based measure where students are asked to read aloud from a passage for 1 minute. The number of words the student read correctly in 1 minute is used as the primary metric to determine student performance (Good et al., 2011b; Shinn, 2008). During benchmark assessments, three ORF passages are administered and the median number of words correct per minute (wcpm) is selected to represent the student's level of oral reading fluency (Good et al., 2011b).

Formative Assessment

Formative assessments are assessment practices designed to routinely monitor and improve learning within the classroom

setting (Berry, 2008). The information gleaned from formative assessments can be used to guide instructional decisions by teachers and to provide students feedback on the progress of their learning. Formative assessments can describe a wide variety of practices ranging from frequent informal observations to more structured curriculum-based measures.

Pennsylvania System of School Assessment

The purpose of the Pennsylvania System of School Assessment (PSSA) is to provide educators, students, parents, and members of the community with detailed information about the about schools performance in meeting the academic needs of the students (Pennsylvania Department of Education, 2010). It is a measure designed to determine the extent to which school curricula help them attain proficiency of academic skills. The reading portion of the PSSA measures five academic skills (Shapiro, Keller, Lutz, Edwards Santoro, & Hintze, 2006). These include learning to read independently, reading critically, reading, analyzing, and interpreting fiction, characteristics and functions of the English language, and research. Students are asked to read a series of passages and answer questions which correspond to these skill areas.

Response to Intervention

Response to Intervention (RtI) refers to a school improvement paradigm that employs a multi-tiered service

delivery model utilizing formative assessments to adjust core and supplemental instruction to ensure positive outcomes for students (Tilly, 2006). It is a framework of instruction used to monitor student academic progress after the implementation of research-based academic or behavioral interventions (Daley, Martens, Barnett, Witt, & Olsen, 2007; Lichtenstein, 2008). It can be employed in various academic areas but has most typically been used for reading, math, and behavior. RtI has been conceptualized in different ways; however, all incarnations contain a few basic components including multi-tier service delivery, the provision of evidence-based interventions, data-based decision making, and formative assessment (Glover & DiPerna, 2007; Kovaleski, 2007).

Specific Learning Disability

According to the Individuals with Disabilities Education Improvement Act (IDEIA, 2004) a specific learning disability (SLD) refers to:

> A disorder in one or more of the basic psychological processes that may manifest itself in the imperfect ability to listen, think, speak, read, write, spell, or to do mathematical calculations, including conditions such as perception disabilities, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia. Such term does not included a learning problem that is

> primarily the result of visual, hearing, or motor disabilities, of mental retardation, of emotional disturbance, or of environmental, cultural, or economic disadvantage(Section §300.307).

Specific learning disabilities are heterogeneous in nature and are not defined by one individual characteristic that represents all individuals with the disorder (Hallahan & Mercer, 2002; Lichtenstein, 2008). The severity of SLD varies from one student to the next and is expressed in different ways depending on the supports available within the learning environment. Students with SLD will likely experience some degree of learning difficulties throughout and the extent of difficulties will depend on the severity of their disorder. Although SLD reflects learning difficulties in a variety of areas, the vast majority of students with SLD have a learning disorder in the area of reading (Lichtenstein, 2008).

Summative Assessment

Summative assessments are assessment practices that gather information regarding the cumulative effect of teaching on student learning (Berry, 2008). The purpose of a summative assessment is to make judgments regarding student achievement against a criterion, standard or norm. Summative assessments can take many forms but are most typically norm-referenced standardized achievement tests, classroom grades and state

proficiency tests. In each case, the goal of the test is to determine what information the student has learned, what skills they are have not mastered, and determine whether or not they have met a preselected standard. Summative assessments are typically not well designed to guide instructional decisions or provide feedback to students regarding their learning.

Universal Screening

Universal screening is a systematic assessment of the students on an academic indicator (Deno, 2003). They are typically assessed three times per year in the fall, winter, and spring of the school year (Hughes & Dexter, 2011). The resultant data from universal screening can be used to highlight the pervasiveness and severity of academic problems (Ikeda, Nessen, & Witt, 2008). Through universal screening data, educators can determine if adjustments are needed to improve the core curriculum and can determine if supplemental interventions beyond the core curriculum are necessary for struggling students. Judgments about the effectiveness of both the core curriculum and interventions supports can be made with universal screening data.

Assumptions

This study is based on several assumptions. First, it is assumed that all assessments utilized, including the DIBELS

and the PSSA, were administered according to standardized procedures. Similarly, it is assumed that the students provided their best effort when completing all assessments. As a review of archival data, steps to ensure accurate assessment procedures were not possible. It is also assumed, and supported by research, that a strong correlation exists between curriculum-based measures and performance on state proficiency assessments thereby ensuring that the benchmark expectations which are derived from the curriculum-based are valid predictors of state assessment performance.

Limitations

The data collected in this study were accessed from two school districts in central Pennsylvania. The demographics represented in both samples may not be reflective of the general population and may not be generalized to other settings. The fall, winter, and spring DIBELS ORF scores were not necessarily collected in both school districts on the same date and in some cases may be more than 3 weeks apart. This could lead to difficulties when interpreting the data. During the data collection period, the students received research-based instruction and interventions designed to accelerate their reading growth. Their resultant progress may have affected the validity of the locally-generated benchmark scores.

This study uses logistic regression to calculate local benchmark scores. This procedure will generate a score which will yield the highest overall predictive accuracy for the sample selected. This generated benchmark, however, may not be generalizable to future cohorts given the potential differences in the demographics of the students and instructional methods and curricula used in the school districts. More research across different settings with samples reflecting a variety of racial and ethnic backgrounds to ensure it can be generalized to other settings.

Summary

This chapter discussed the application of curriculum based-measurement benchmark expectations. Background for the rationale and application for benchmarks expectations was discussed. Information on the relationship between curriculum-based measurement benchmarks and performance on state-proficiency was presented. Problems associated with national normed benchmark expectations were considered and a rationale for the use of locally-generated benchmark expectations was explored. Finally, research questions, definitions of terms, assumptions and limitations were identified and discussed.

CHAPTER II

REVIEW OF THE RELATED LITERATURE

This chapter will discuss the literature related to the development of locally-generated benchmark scores. As a framework for the interpretation of the benchmarks, the core elements of a response to intervention model (RtI) including multi-tier service delivery, provision of evidence-based interventions, formative assessment, and data-based decision making are discussed. A particular focus on the assessment procedures and decision making practices utilized in the RtI process is provided to ensure a strong research-supported foundation exists on which reliable and valid benchmark scores can be constructed. The findings of the studies included in this review show strong correlations between curriculum-based measurement (CBM) in reading and performance on both standardized measures of reading achievement and state test performance. Similar correlations were also demonstrated between CBM and the scores Pennsylvania System of School Assessment (PSSA). As a result of these strong relationships, the extension of CBM in reading to predict performance on state assessments through the use of locally-generated benchmarks appears to be supported by research evidence. Concluding this chapter is a discussion of the characteristics of the assessments used in this study including both the

Dynamic Indicators of Basic Early Literacy Skills (DIBELS) and the PSSA. In addition, the procedures for calculating benchmark expectations provided by the DIBELS System along are provided. This is followed by a rationale for utilizing locally-generated benchmarks. Finally, the method used for calculating the locally-generated benchmarks and a rationale for why this method was chosen are discussed.

Identification of Learning Disabilities

Ability/Achievement Discrepancy

The ability/achievement discrepancy approach to the identification of specific learning disabilities (SLD) began in 1977 with the passage of the final regulations of the Education for All Handicapped Children Act (EAHCA; Lichtenstein, 2008; Meyer, 2000). The criteria indicated that a student with a learning disability is one who does not achieve at a level commensurate with age and ability level and who has a severe discrepancy between ability and achievement in one or more academic areas (Hallahan & Mercer, 2002).

Unfortunately, the validity of the ability/achievement criteria has been called into question. Supporting the criticisms are research findings are that undermine the basic tenants of the ability/achievement discrepancy criteria. A specific learning disability has been historically defined as an unexpected underachievement indicating that a student is

learning disabled when they, in the absence of other causal factors, do not achieve at a level commensurate with their level of ability (Lichtenstein, 2008; Rutter & Yule, 1975).

Two problems are apparent with this conceptualization of SLD. First, a perfect correlation does not exist between intelligence and achievement. According to Sattler (2001), correlations between intelligence measures and tests of achievement are rarely above .60 accounting for 36% or less of the shared variance. With this disconnect between the assessment measures, the application of the ability/achievement criteria would yield far too many false positives and false negatives to be used in such an important role (Lichtenstein, 2008). Also, to validate the significance of an unexpected underachievement, these criteria must confirm that students with ability/achievement discrepancies suffer from a unique condition that is separate from other subtypes of struggling learners, particularly students with low intellectual ability. Unfortunately, support for this dichotomy has not been shown in the research literature. Substantial differences between the skill deficits of students with ability/achievement discrepancies and students with low-ability cannot be found (Hallahan & Mercer, 2002; Lichtenstein, 2008). The learning problems that are at the root of SLD impede the acquisition of skills in the same way

for both high-ability and low-ability struggling students (Lichtenstein, 2008).

The ability/achievement discrepancy criteria has also been criticized as a wait-to-fail model (Hallahan, & Mercer, 2002; Lichtenstein, 2008; Speece, 2002). Students having academic difficulties in the early grades will likely not show the discrepancy necessary for SLD eligibility until the later grades (Restori, Katz, & Lee, 2002). This is because the norm-referenced standardized achievement tests used in the discrepancy model are not sufficiently sensitive to detect deficient skills in the early grades. To obtain an average or non-discrepant standard score, a student may need to correctly answer one to two items correctly on a subtest. On a number of standardized achievement tests, even a score of zero will not return a significantly below average result. In the absence of sufficiently sensitive assessments, students must wait until the latter grades for a discrepancy to appear in order to receive special education supports. This is an unfortunate consequence given that researchers have indicated that intervention efforts are more effective the earlier they are put into action (Hallahan & Mercer, 2002).

Another important concern is that students showing long term difficulties but who do not demonstrate an ability/achievement discrepancy may be denied special

education supports (Restori et al., 2002). These low ability readers may show academic skills that are equivalent or below those of their learning disabled peers. Since the student's general intelligence is low, the poor achievement is expected and not indicative of a learning disability. This is particularly problematic when the student's level of ability is not low enough to qualify as a student with an intellectual disability (Lichtenstein, 2008). An unfortunate dilemma is created for school psychologists. They are forced to choose between inaccurately qualifying a non-discrepant student with low intellectual ability as learning disabled or must allow the student to proceed in the regular education setting without the services they need to succeed (Restori et al., 2002).

As the criticisms mounted, alternatives to the ability/achievement discrepancy approach to SLD identification were explored (Hallahan & Mercer, 2002; Lichtenstein, 2008). After the passage of The Individuals with Disabilities Education Improvement Act (IDEIA; 2004) school districts are no longer required to use a severe discrepancy between ability and achievement to determine whether a child has a specific learning disability. Although, the continued use of the ability/achievement discrepancy is permitted the use alternative research-based procedures or a process centered on

a student's "response to scientific, research-based interventions" to determine the presence of a learning disability are provided for within IDEIA (Section §300.307). These alternatives fall into one of two general and philosophically opposing paradigms; cognitive process approach and a response to intervention (RtI) model.

Cognitive Processes Approach

Wording in the Federal definition of a learning disability includes the phrase "a disorder in one or more of the basic psychological processes" (Section §300.8). Cognitive processes researchers have latched on to this phrase and have developed a system of assessment to identify the presence of SLD (Flanagan, Fiorello, & Ortiz, 2010; Hale, Kaufman, Naglieri, & Kavale, 2006). Rather than simply evaluating the difference between general intellectual ability and academic achievement, this approach evaluates the intrinsic cognitive weaknesses that are the foundation of problems in academic performance (Torgesen, 2002). According to Hale et al. (2006) the evaluation of intrinsic cognitive weaknesses is conducted to uncover what is thought to be the primary feature of SLD "a consistency between cognitive deficits and academic deficits coupled with a significant discrepancy between cognitive strengths and cognitive deficits" (p. 357).

Standardized cognitive assessments are at the heart of a cognitive processing approach (Hale et al., 2006). The assessments employed in this approach are aligned with contemporary models of learning, cognition, and intelligence. Hale and Fiorello (2004) propose that several assessments have been identified as appropriate for use in this type of model as they have strong theoretical foundation. These assessments may include but are not limited to: the Differential Ability Scales (DAS; Elliot, 1990), the Stanford Binet Intelligence Scales: Fifth Edition (Roid, 2003), the Wechsler Intelligence Scale for Children – Fourth Edition (Wechsler, 2003), and the Woodcock-Johnson Tests of Cognitive Abilities – Third Edition (Woodcock et al., 2001). Each of these measures assesses varying aspects of cognition and intelligence and their use are dependent on the referral question asked (Hale & Fiorello, 2004). The evaluator in this model may have to use subtests from multiple instruments in order to comprehensively assess the cognitive processes relevant to the academic deficit.

Three main models have been proposed for evaluation of cognitive processes. Naglieri (as cited in Hale et al., 2006) proposed the *Discrepancy/Consistency* model this approach evaluates student performance with the goal of identifying inter- and intra-cognitive weakness (Flanagan, Fiorello, & Ortiz, 2010). This model sets its foundation in the Planning,

Attention, Simultaneous, and Successive (PASS; Das, Naglieri, & Kirby, 1994) theory of intelligence. In this model, the identified weaknesses must be consistent with academic weakness. In addition, students must show particular cognitive processes strengths that are significantly better developed than the cognitive weaknesses.

A second cognitive processes assessment model was proposed by Flanagan, Ortiz, Alfonso, and Mascolo (as cited in Flanagan, et al., 2010) referred to as the *Operational Definition* approach. According to Flanagan, Fiorello, and Ortiz (2010) this method three levels of evaluation to uncover and understand the relationships between cognitive abilities and academic weaknesses. In addition, exclusionary factors are evaluated to distinguish between SLD and academic disorders caused by other disabilities (e.g. intellectual disability, emotional disturbance, behavior problems, and language disorders). SLD is then evaluated within the operational definition of the disorder that states "below average aptitude-achievement consistency within an otherwise normal or average ability profile" (p. 742). Aptitude-achievement consistency is uncovered according to the literature related to the Catell-Horn-Carrol (CHC; Horn & Catell, 1967) theory of intelligence.

A third cognitive assessment model, Hale and Fiorello's (2004) Cognitive Hypothesis Testing (CHT), is based on four premises. First, cognitive process are have been linked by empirical evidence to achievement in reading. Second, students often demonstrate a unique pattern of strengths and weaknesses. Third, these unique profiles of cognitive processes must be directly assessed within the context of ecological and treatment validity. Finally, interventions must be designed to remediate and accommodate for the cognitive deficits that underlie the reading disorder. The key element of this approach is the relationship or concordance that is present between cognitive process and academic deficits (Flanagan, Fiorello, & Ortiz, 2010). Students must also have cognitive strengths that and there is inconsistency or discordance between the strengths and the academic skill deficit. This model is grounded in the CHC theory of intelligence but is additionally supported through a neuropsychological understanding of the underpinnings of an academic disorder (Hale & Fiorello, 2004).

Regardless of the specific method employed, the basic tenant of a cognitive processes approach measures the connections between norm-referenced standardized assessments of intellectual ability and academic achievement (Hale & Fiorello, 2004; Lichtenstein, 2008; Stuebing, Fletcher,

Branum-Martin & Francis, 2012). Hale and Fiorello (2004) suggest that the evaluator seeking to identify a disorder a successful evaluator must understand the intricacies of the cognitive assessment, be able to identify the relationships between the subtests, and recognize the brain functions to which they relate. This will allow the evaluator to link the cognitive strengths and weaknesses to academic difficulties and, ultimately, to appropriate and effective academic interventions (Lichtenstein, 2008; Stuebing et al., 2012). For example, several cognitive deficits have been linked to weaknesses in reading achievement including auditory processing, short-term memory, long-term memory and retrieval, processing speed, crystallized intelligence (Hale et al.,2006; Hale & Fiorello, 2004). The evaluator may find that a student with significant difficulties in reading will also show deficits in one or more of these cognitive abilities. The resultant findings are analyzed in combination with classroom-based assessments and observations, to formulate a diagnostic impression of whether a SLD is present (Hale & Fiorello, 2004).

Cognitive processing methods for SLD identification have several drawbacks that cause researchers to question their use. First, there is not enough research support to guide decisions regarding the examination of cognitive strengths and

weaknesses to uncover the presence of SLD. School psychologists must rely heavily on professional judgment to make connections between cognition and achievement and draw appropriate conclusions with their findings (Fletcher, Coulter, Reschly & Vaughn, 2004). Another concern with this approach is that learning problems are highly dependent other factors that are independent of cognitive processes (Lichtenstein, 2008). These factors may include the learning environment, availability of interventions supports, positive reinforcement cultural and environmental influences and the quality of the teacher. Significant effort must be undertaken to rule out these other causal factors in order to make an accurate identification of SLD.

A third major issue with cognitive processing approaches to SLD relates to the relationship between the evaluation and recommendation. A long history of research findings has shown that meaningful connections cannot be made between the results of these assessment and the resultant recommendations and interventions supports that are developed (Lichtenstein, 2008; Melby-Lervåg & Hulme, 2012; Shipstead & Redick, 2012). For example, Melby-Lervåg and Hulme (2012) conducted a meta-analysis of 30 studies to determine if working memory training programs improve other cognitive abilities as well as academic achievement in both children and adults. The authors found

that these programs consistently produced short-term improvements with verbal and nonverbal memory tasks. These gains, however, were not sustained during follow-up evaluations assessed an average of 9 months later. More importantly, the skills learned through these programs do not appear to generalize to gains in other areas such as verbal ability, arithmetic, or word decoding. Based on their findings, the authors suggest that working memory procedures cannot be recommended as appropriate interventions for the treatment of developmental disorders, including dyslexia.

According to Bradley, Danielson, & Hallahan (2002), the available methods for assessing deficits in cognition are insufficient. Consequently, measuring cognitive processes in an attempt to connect them with deficits in achievement and ultimately to interventions is not practicable. As the theory and assessments used in this approach improve, the use of this alternative may become more widely used. Until this time, other methods of SLD identification may be more appropriate. Bradley, et al. (2002) further suggest that:

> Processing deficits should be eliminated from the criteria for classification because no clear measure or understanding of processing deficits currently exists. Although evidence exists that individual with SLD have processing limitation, methods for measuring the presence

of processing difficulties and devising appropriate interventions for those deficits have yet to be established (p. 797).

Response to Intervention

The second alternative to the ability/achievement discrepancy criteria, which will remain the focus of this dissertation, is the Response to Intervention model. Response to Intervention (RtI) is a framework of instruction used to monitor student academic progress after the implementation of research-based academic or behavioral interventions (Daley, Martens, Barnett, Witt, & Olsen, 2007; Lichtenstein, 2008). It can be employed in various academic areas but has most typically been used for reading, math, and behavior. RtI has been conceptualized in different ways; however, all incarnations contain a few basic components including multi-tier service delivery, the provision of evidence-based interventions, formative assessment, and data-based decision making (Glover & DiPerna, 2007; Kovaleski, 2007).

Multi-tier service delivery. Services in RtI are provided through a multi-tier framework which is almost universally conceptualized as a pyramid (Tilly, 2008). The number of tiers in an RtI model can vary and may include several levels. Most RtI models, however, follow a three tiered approach. At Tier 1, all students are provided

instruction within in the general education setting (Daly, Martens, Barnett, Witt, & Olsen, 2007; Glover & DiPerna, 2007; Hughes & Dexter, 2011; Tilly, 2006). According to Tilly (2008) the majority of schools have implemented a reading block of 90 minutes of instruction per day using the Tier 1 core curriculum. During this block of time, all students are taught using the research-based curriculum selected by the school district. Distractions are kept to a minimum and teachers are encouraged to instruct using the core curriculum with a high degree of fidelity. This is because the use of scientifically-based instruction that uses effective teaching principles will likely result in the majority of students progressing as expected in reading acquisition (Hughes & Dexter, 2011; Tilly, 2008).

The scientifically-based core curriculum applied in Tier 1 refers to instructional programs and practices that are aligned to research-supported conceptions of development that are designed to meet the needs of the majority of students (Hughes & Dexter, 2011). Within reading, the findings of the National Reading Panel (NRP; NICHD, 2000) create a strong foundation for a research-supported core curriculum. Through a meta-analysis of over 100,000 studies, the NRP identified five components of effective early reading instruction (McCardle & Chhabra, 2004). These Five Big Ideas include:

phonemic awareness, phonics, fluency, vocabulary, and text comprehension. The inclusion of instructional practices and techniques derived from these five big ideas are an important first step in development of a research-supported core-curriculum for reading instruction.

The implementation of a scientifically-based instruction that is implemented with integrity is expected to result in positive outcomes for 80 - 90% of all students (Shapiro, 2008; Tilly, 2008). This would indicate that in a classroom of 30 students, 24 – 27 should progress without the need for significant supplemental interventions. This percentage is an estimate based on the research available. According to Tilly (2008), "it is a logical and rational approximation of how effective core instruction should be" (p. 31). If a high percentage of students require supplemental instruction, the school system may lack the resources necessary to provide supplemental supports. Schools with a high percentage of struggling students would be better served by allocating more resources to improving the core instruction rather implementing supplementary interventions.

Unfortunately, not all students will successfully respond to the Tier 1 core curriculum alone. Based on the results of formative universal screening assessments, a minority of students will be identified as in need of targeted

intervention supports (Gresham, 2008). These Tier 2 research-based interventions are provided to students identified as at-risk in addition to the instruction within the core classroom curriculum. Tier 2 interventions are designed to be very efficient and produce relatively quick results (Daly et al., 2007; Glover & DiPerna, 2007; Gresham, 2008; Hughes & Dexter, 2011; Tilly, 2006). Ideally, no more than 20% of students will be in need of a Tier 2 interventions (Dexter & Hughes, 2007; Gresham, 2008). The amount of time a student will receive a supplemental Tier 2 intervention may vary depending on the length of time required by an intervention. A minimum of 30 minutes of Tier 2 instruction is typically recommended in addition to the 90 minutes of core instruction (Dexter & Hughes, 2007).

At the apex of the RtI pyramid, increasingly intensive interventions are delivered to support struggling students. Based on the results of ongoing progress monitoring, it is possible that some students will fail to demonstrate adequate response as a result of the Tier 2 interventions and universal core instruction at Tier 1 (Gresham, 2008). For these students more intensive Tier 3 interventions are provided in addition to the core curriculum instruction and Tier 2 intervention (Glover & DiPerna, 2007; Gresham, 2008; Hughes & Dexter, 2011; Tilly, 2006). This level of support is reserved

for students that show long-term academic difficulties (Gresham, 2008). Interventions at this level must be comprehensive and intensive and will likely need to be implemented over a long period of time. Ideally, no more that 5 - 10% of students should require a Tier 3 intervention (Dexter & Hughes, 2007; Gresham, 2008).

Provision of evidence-based interventions. At each of the three levels of the RtI pyramid, evidence-based interventions set the foundation for sound instructional practices (Glover & DiPerna, 2007). This firm grounding will help to ensure that students receive maximum benefit from RtI services. In addition, instruction that has been validated by research helps to identify effective practices and programs and highlights essential information on why they work (Reyna, 2004).

Universal or Tier 1 interventions are designed to meet the needs of all students (Gresham, 2008). They are given to all students in the same way within the same environment and are typically not individualized. These interventions are provided to prevent student difficulties before they occur. Within reading, universal interventions may include the school district's reading curriculum but also may include effective instructional strategies such as modeling, prompting and error correction, opportunities to respond, and shaping and

reinforcement strategies (Burns, VanDerHayden, & Boice, 2008; Gresham, 2008).

Targeted or Tier 2 interventions focus on students who do not show adequate response to the Tier 1 or universal intervention. They are designed to focus more on the individual needs of a student identified as at-risk. These interventions should be efficient, robust and produce results in a relatively short period of time (Gresham, 2008). Tier 2 interventions are typically delivered in a small group setting or within the regular classroom during a differentiated instruction lesson by the teacher or instructed by support personnel. In reading, the provision of targeted interventions are dependent on the skill deficit present. For example, a student demonstrating a deficit in reading fluency may benefit from a repeated reading or incremental rehearsal strategy (Burns, VanDerHayden, & Boice, 2008). In addition, the provision of a standard protocol intervention that has standardized instructions and specific scope and sequence may be provided at this level. For a student with difficulties in reading fluency, she or he may receive instruction using the Read Naturally (Inholt, 1991) standard protocol intervention that is designed specifically to remediate reading fluency deficits.

Intensive or Tier 3 interventions are provided to students who have not responded adequately to prior instructional efforts and will need the most intensive level of support (Gresham, 2008). Students receiving intensive interventions show long-term deficits that are resistant to typical instructional practices. Intensive interventions should be highly individualized and may require a comprehensive evaluation of performance and in some cases an Individualized Education Program (IEP). The intervention provided should be systematic, intense, and provided over an extend period of time. Standard protocol interventions also may be used as a Tier 3 intervention. The Corrective Reading program (Engelmann et al., 1999) and the Wilson Reading System (Wilson, 1988) are two standard protocol reading interventions that are typically used at Tier 3.

At each level it is important to evaluate the level of integrity at which the intervention was employed (Kovaleski, 2007). A high level of treatment integrity has been shown to increase academic achievement (Upah, 2008). In addition, interventions that have been employed with a high degree of fidelity will allow for sound data based decisions to be made about student performance. If standardized procedures are not followed then drawing reasonable conclusions about student progress becomes more difficult. Conversely, it becomes

nearly impossible to attribute success or failure to the intervention if it is not carried out as designed (Glover & DiPerna, 2007; Upah, 2008). To ensure fidelity of the intervention, several structures should be put in place including, self-report measures, ongoing professional development, peer supports, modeling practices, and structured observations (Upah, 2008).

Formative assessment. A formative assessment measures student achievement to determine if instructional practices are effective and which ineffective strategies should be modified or eliminated (Shinn, Shinn, Hamilton, & Clarke, 2002). Three levels of formative assessments are employed in an RtI model including universal screening, strategic monitoring, and frequent progress monitoring (Hasbrouck & Tindal, 2007). These levels of assessment serve different functions and their application will vary depending on the individual needs of the student.

Universal screening measures correspond to Tier 1 of the RtI pyramid and are the initial step to identify students in need of targeted interventions (Hughes & Dexter, 2011). Through universal screening, all students are assessed three times per year in the fall, winter, and spring of the academic school year (Hughes & Dexter, 2011; Shinn, 2008, Tilly, 2008). After screening, the collected data are analyzed. Based on

the disseminated information, students who are likely at-risk are identified and provided with additional supports (Ikeda, Neessen & Witt, 2008).

Ideally, universal screening assessments are cost-effective, quick and easy to administer tasks that are highly predictive of overall academic outcomes (Deno, 2003; Ikeda, Neessen & Witt, 2008). According to Kovaleski and Pedersen (2008), universal screening measures should meet several characteristics. First, the assessments should be closely linked to state or national standards. Within reading, assessment procedures that that test elements of the five big ideas in reading (NICHD, 2000) should be particularly useful if the core reading curriculum employed by the school district is supported by research. Second, universal screening should be able to be administered to large groups of students in an efficient manner. This ensures that time needed to complete the assessments does not interfere with instructional practices. Third, these measures must be administered frequently and sensitive to small amounts of change so that student growth can be effectively assessed. This allows for valid decisions to be made regarding student skill development within a short period of time. Educators will not have to wait until the results of outcome measures like state proficiency to have a picture of how students are progressing.

Finally, the resultant data must be user-friendly and efficiently organized to show student progress on each individual skill assessed. This allows educators from a variety of backgrounds with varying levels of expertise to interpret and apply the data.

CBM procedures are one such assessment practice which meets all of the criteria of an effective universal screening measure (Kovaleski & Pedersen, 2008). It is a group of brief standardized assessments used by educators to measure the impact of their instructional practices (Shinn, 2008). They are typically fluency-based measures where students are expected to demonstrate the desired academic skill for only a short period of time, approximately 1 - 3 minutes. In reading, two CBM have primarily been employed for the purpose of universal screening. These include Oral Reading Fluency (ORF) and Maze assessments.

More frequent monitoring may be used to ensure an adequate trajectory of growth is made for students receiving Tier 2 interventions (Shinn, 2008). This allows for ongoing curriculum adjustments to be made if the student is not growing as expected. These assessments may repeat a version of the CBM used for universal screening more frequently, usually 2 - 3 times per month. The increased frequency allows for minor changes to the intervention provided at Tier 2 to be

made. In addition, more data can be gathered using frequent monitoring to determine if more or less intensive interventions are needed (Hasbrouck & Tindal, 2006; Ysseldyke, et al., 2010).

Students deemed non-responsive at Tier 2 and who receive Tier III interventions may require an even more frequent level of formative assessment. Progress monitoring at this level increases the frequency of the assessments used at Tier 2 to 1 – 2 times per week (Hasbrouck & Tindal, 2006; Shinn, 2008; Ysseldyke, et al., 2010). This level of monitoring is done to ensure that effective interventions are provided and ineffective practices can be changed as soon as possible (Shinn, 2008). A majority of students who receive this level of monitoring may receive special education supports or could receive special education in the near future if Tier 3 intervention supports are unsuccessful.

Decision making in an RtI framework. Kovaleski and Petersen (2008) suggest that collaborative teams should be employed at all levels of the RtI pyramid to ensure high-quality instructional decisions are made. These problem-solving or Data Analysis Teams provide a structured format that allows team members to decide what interventions will be provided, how long they will be implemented, and what level of support is necessary. To make these decisions, the Data

Analysis Team analyzes CBM universal screening data to distinguish the students who are responding to the core curriculum from those who are in need of additional supports. This decision may be based on a benchmark or normative approach where students who fall below a selected percentile criterion (e.g., 25th percentile) or benchmark are identified as at-risk and will receive Tier 2 supports (Ysseldyke et al., 2010). In reading the DIBELS provide an easily interpretable set of benchmark expectations which identify students who are *Low-Risk*, *Some-risk*, or *At-risk*. After universal screening with DIBELS, DATs can identify which of the three categories students fall and allocate the intervention based on the information.

At Tiers 2 and 3 the Data Analysis Team analyzes progress monitoring data to determine if the intervention is meeting the needs of the student (Burns, Wiley, & Viglietta, 2008). The team evaluates the rate of progress made and determines whether the student progress indicates the need for further intervention or a reduction of services. In addition, decisions about increasing the level of intervention from Tier 2 to Tier 3 or from Tier 3 to special education made.

Students who are unable to demonstrate adequate response to the interventions provided in Tiers 1 - 3 may be found eligible for a special education placement as a student with a

specific learning disability (Shapiro, 2008; Tilly, 2008). To determine a student's response to an intervention and potential special education eligibility, researchers have recommended a duel discrepancy model using CBM (Fuchs & Fuchs, 1998). This model evaluates two important features of achievement: level of performance compared with same aged peers and rate of learning (Fuchs, 2003; Fuchs & Fuchs, 2007; Speech, 2003; Speech & Case, 2001). Determination of insufficient rate of improvement (ROI) is calculated by measuring student progress on reliable measures over a sufficient period of time (Fuchs & Fuchs, 2007; Fuchs, Fuchs, Hamlett, Waltz, & Germann, 1993). Insufficient growth may be observed as a 1 standard deviation difference between the student's rate of improvement and that of same-age peers (Fuchs & Fuchs, 1997). In addition, a normative approach that rank orders the growth rates of student CBM data and subsequently compares students against percentile thresholds, has also been proposed to identify at-risk students (Burns & Senesac, 2005).

The level of performance in a dual-discrepancy approach is evaluated through a comparison between current level of functioning and the level of functioning of same-age peers in relation to identified standards (Silbergiltt & Hintze, 2007). Some disagreement exists on how to best determine insufficient

level of performance (Burns & Senesac, 2005). Performances on a norm-referenced standardized achievement test or the results of state assessment may provide an accurate measure of deficient level of performance (Torgesen et al., 2001). For example, normative approaches using CBM percentile scores and standard scores on a nationally normed standardized measure of achievement have also been presented as viable options (Fuchs, 2003; Torgesen, Alexander, Wagner, Rashotte, Voeller, & Conway, 2001). Gresham (2008) suggests that performance below the 10th percentile on a CBM should warrant further evaluation. The use of student scores in relation to CBM benchmark expectations has also been presented as an appropriate method (Burns, 2008; Gresham, 2008). Gresham (2008) suggests that a 50% discrepancy between the student's score on a CBM and the normative expectation for the grade level may indicate the need for further evaluation. Similarly, Burns (2008) illustrated the use of locally-generated benchmark expectations to create a comparison for the level of performance in a dual-discrepancy approach. Universal CBM like DIBELS establish criterion-referenced benchmarks to help identify struggling students in dual-discrepancy approach (Ardoin & Christ 2008; Hughes & Jenkins, 2011). Through this setup educators can measure students' level of performance compared to benchmarks expected for children at their grade

level and make reliable decisions regarding the intervention needs of the students (Burns, 2008).

Fuchs and Fuchs (2003) suggest that dual-discrepancy model is based on three assumptions. First, student's abilities are varied. Each student will receive different educational experiences even within the same classroom setting. Second, academic difficulties are a function of the educational environments in which the student resides. If students with low academic skills are progressing at rate consistent with same-age peers, then a learning disability is not present. Third, if most of the students in a classroom are showing inadequate growth, the enhancement of the classroom instruction for all students should be the primary focus before considering a learning disability for one student.

Advantages and research needs of RtI. The RtI model has several advantages that make it an appealing framework for the provision of regular education supports and as an alternative for SLD identification. First and foremost, RtI has strong focus on the prevention of learning difficulties (Fuchs & Fuchs, 2007). In this model, students receive ongoing evidence-based supports and interventions that are designed to specifically address their needs. These supports will lead to enhancements in instruction and student improvement within the

general education setting when they are implemented with fidelity (Lichtenstein, 2008). In addition, the use of CBM in RtI allows for reliable and valid assessments that are strongly connected to instructional practices (Fletcher, et al., 2004; Glover & DiPerna, 2007; Lichtenstein, 2008). As students progress through the tiers, ongoing formative assessments can be used to adjust the classroom instruction or supplemental intervention to ensure a match is made between student and curriculum.

The ongoing formative assessments and intervention supports are what make the RtI model a viable alternative for the identification of SLD. Through these features, the *unexpected underachievement* concept that is at the center of the disability is addressed (Fletcher et al., 2004; Lichtenstein, 2008). According to Fletcher et al. "a student with LD is identified as one who has unexpected difficulty learning and the discrepancy is measured relative to the expectation that most students can learn if quality instruction is provided" (p. 313). Similarly, RtI avoids the wait to fail phenomenon that is the result of the ability/achievement criteria (Fletcher, et al., 2004). Rather than waiting for the results from a protracted and often delayed evaluation, students are provided ongoing supports and accommodations without delay. Consequently, eligibility is

not derived from isolated test scores but from a system of instructional enhancements and ongoing progress monitoring (Fletcher et. al, 2004) When students fail to show significant progress Tiers 1 - 3, their lack of response as measured by formative assessments becomes the reliable and valid data that can be used for the identification of SLD (Fletcher, et al. 2004).

Several future research literature needs of the RtI were identified by Glover and DiPerna (2007). First, the individual and group outcomes of tiered interventions need further evaluation to ensure quality and to determine if the interventions are indeed remediating the deficits of the students who receive them. Similarly, the effects of individualizing intervention components and intensity should be further examined. More studies need to be conducted that identify the critical elements within interventions and sort out the elements that improve responsiveness for students. Another important need is the constant evaluation of the many intervention approaches used in RtI model. The provisions in NCLB (2000) and IDEIA (2004) make a number of references for the use of evidence-based interventions. As a result, a proliferation of supplementary programs is available for purchase from a large number of publishing companies. Many of these publishers claim that their products are research-

supported. Unfortunately, these claims are not always backed by reliable research findings. Ongoing research must be conducted to ensure the quality of all interventions used within an RtI framework.

Steps also need to be put in place to improve the measurement procedures and decision-making accuracy used in the RtI process (Glover & DiPerna, 2007). First, more information is needed that evaluates the compatibility and utility of various assessment tools and data-based criteria. Although a great deal of research is available measuring the reliability and validity of assessment practices used in RtI, more research should be directed toward the use of CBM as it applies to data-based decision for movement along the tiers. In addition, future research and development are needed to determine the utility of decision-making criteria for both the selection of interventions and the determination of student responsiveness. More research is needed to ensure that the balance of sensitivity, specificity, negative predictive power (NPP), and positive predictive power (PPP) for CBM cut-scores are acceptable for the determination of students as at-risk. This will help to reduce the number of false positives and inaccurate classifications inherent in universal screening procedures. As with any measure, a continuing evaluation of

the psychometric integrity and use of CBM across diverse populations and contexts must be conducted.

The primary objective of this study should address a number of the suggestions for future research given by Glover and DiPerna (2007). The decision-making criteria through the RtI model will be expanded to include the formulation and interpretation of locally-generated CBM benchmarks. Once developed, these benchmarks may improve universal screening procedures and assist special education eligibility determination. Since they are developed from a local sample, representative of the population where they are derived, they should produce a reliable set of standards that are consistent with local expectations. This local connection between the benchmark expectations and student population will be applied with the goal of improving the accuracy of universal screening classification decisions as well as the accuracy of special education determination within an RtI model. The following section further addresses this goal through a discussion of the instrumentation used to develop the benchmark scores.

Instrumentation

The purpose of this dissertation is to analyze the connection between the ORF benchmark expectations and performance on state proficiency assessments. To analyze this connection, the relationship between curriculum-based measures

and tests of overall reading achievement must first be analyzed. Correlations between CBM in reading, particularly ORF, and measures of general reading achievement as well as state test performance are presented. This review is done to ensure a strong research-supported foundation exists on which reliable and valid benchmark scores can be constructed.

Curriculum-Based Measurement

CBM is a group of brief standardized formative assessments used by educators to measure the impact of their instructional practices (Shinn, 2008). Originally developed as a tool to assist special education teachers in the implementation and monitoring of Individualized Education Program (IEP) goals, its use has since expanded to the regular education setting (Deno, 1985, 2003). Deno (2003) suggests several common uses of CBM. First, CBM can be used to improve instructional programs. This can be employed at all three levels of the RtI pyramid. At Tier 1, universal screening data using CBM can determine which students are in need of additional supports. At Tiers 2 and 3 CBM progress monitoring data can be used to determine the amount of growth students have made in response the interventions they received. In addition, judgments can be made about the effectiveness of the instruction that the students receive or determine if a match between student and instruction has been made. A second

important use of CBM is the prediction of performance on important outcomes. Data collected from CBM can be used to accurately show which students will meet proficiency and those who will not. The ease of communication between parents, teachers and students can also be increased with CBM. Through the uses of graphs and easily understandable data, interested individuals can interpret CBM data with little or no difficulty. This is particularly useful within Data Analysis Teams (Kovaleski & Pedersen, 2008) and Problem Solving Teams (Burns, Wiley, & Viglietta, 2008) where individuals from disparate educational discipline coordinate to determine student instructional needs. CBM also helps to reduce the bias in assessment as it decreases the subjectivity of the judgments made by teachers and assessment professionals. It is highly useful to assist instructional planning to help educators determine what instructional practices are successful and what skills need to be taught.

CBM procedures are reliable and valid, simple and efficient to administer, able to be frequently assessed with results that are easy to communicate (Deno, 1985). They are typically fluency-based measures where students are expected to demonstrate the desired academic skill for only a short period of time, approximately 1 - 3 minutes (Shinn, 2008). CBM can be used to measure a wide variety of skills; however,

CBM typically assesses ORF, maze, spelling, written expression, math computation, and applications. Although CBM measures can be developed within a school district, a number of assessment packages are commercially available (Kovaleski & Pedersen, 2008). Two of the most popular CBM systems are the AIMSweb system (Shinn & Garman, 2006) and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good et al., 2011a). Both systems provide a user-friendly structure for the assessment and analysis of CBM.

CBM as a formative assessment is designed to continually monitor student progress to inform instructional practices to improve student learning (Berry, 1998; Deno, 2003). In the present study, CBM is extended beyond this formative role. It is applied as standard that can be used to predict performance on a summative measure of student achievement. This application of CBM may be invalid if a substantial relationship between formative and summative assessments is not present. In the following sections, this relationship is analyzed in further detail. Twenty studies are presented that evaluate the relationship between CBM and performance on summative assessments. Various versions of summative assessments are presented that include both norm-referenced standardized achievement test of differing types and several state assessments. The goal of this analysis is to

demonstrate a strong relationship between the formative CBM and summative assessments. By extension, this will provide an argument for a strong relationship between ORF and general reading achievement. A strong relationship between these concepts will provide a sturdy theoretical foundation on which the locally-generated benchmark scores can be constructed.

CBM and reading achievement. A great deal of research has validated the use of CBM as a reliable and valid measure of general academic outcomes (Shinn, 2008). CBM has been shown effective for instructional programming, improving educational programs by incorporating target goals, formative assessments, and program evaluations (Hintze & Silberglitt, 2005). CBM have shown to be reliable and valid predictor of reading performance across a wide variety of settings and various student characteristics including race, sex, and socio-economic status (Baker et al., 2008; Baker & Good, 1995; Dominguez de Ramirez & Shapiro, 2006; Paleologos & Brabham, 2011; Pearce & Gayle, 2009; Wiley & Deno, 2005; Wise et al., 2010).

Deno, Mirkin, and Chiang (1982) conducted three studies to examine the concurrent validity of curriculum-based reading measures with standardized measures of reading achievement. Three standardized achievement measures and five formative reading measures were analyzed. In the first study, 18

regular education students and 15 students with learning disabilities were randomly selected in grades 1 - 3 from a Minnesota public school. Each student was assessed with the 5 formative assessments including words in isolation, words in context, oral reading, cloze comprehension, and word meaning. In addition they were assessed with the Reading Comprehension subtests the Stanford Diagnostic Reading Test and the Reading Comprehension and Word Identification subtests from the Woodcock Reading Mastery Test. Correlation coefficients between the reading aloud measures (words in isolation, words in context, and oral reading) and the three criterion measures ranged from .73 - .83. Correlations between the cloze and word meaning indicators ranged from .60 - .83. Negligible differences were uncovered between correlations identified for the regular education and special education groups for the reading aloud measures. This was not true for the cloze and word meaning measures as correlations for the special education group were lower and ranged from .59 - .83.

In the second study, 27 regular education students and 18 students with learning disabilities in grades 1 – 6 were randomly selected from two Minnesota public schools (Deno et al., 1982). This study was conducted to determine if the grade level of the material or the duration of the test alters the correlations. They found that correlations between the

third- and sixth-grade level materials were in the .80 - .90 range. Tests assessed for 30 seconds were highly correlated with 1-minute tests with correlations consistently greater than .90.

To replicate the findings of the first two studies, the authors randomly selected 43 regular education and 23 students in grades 1 - 6 from three inner-city schools in Minnesota (Deno et al., 1982). Each student was assessed with three formative measures, words in isolation, oral reading, and a cloze reading passage. The participants were also assessed with the Phonetic Analysis and Reading Comprehension subtests from the SDRT and the Reading Comprehension subtest from the Peabody Individual Achievement Test (PIAT). Intercorrelations between the formative measures were high according to the authors (coefficients were not reported). Correlations between oral reading and the reading comprehension subtests were particularly strong ranging from .78 - .90 and were stronger than the correlations obtained between the cloze comprehension and the reading comprehension subtests (.67 - .80).

Upon completion of the three studies, Deno et al. (1982) make three important conclusions. First, measures which assess students by asking them to read aloud closely relate to performance on standardized reading assessments. Second,

measuring correct performance is more valid approach than assessing error performance. They recommend assessing the number of items answered correctly rather than the number of mistakes made. Third, the authors found that to obtain a strong relationship with a standardized measure of reading comprehension, a student only needs to read aloud for 1 minute. In actuality, they determined that 30 seconds is sufficient to obtain a valid indicator of student skills; however, they recognize that formative measures are typically assessed in 1 minute segments.

Jenkins and Jewell (1993) also compared the relationship between standardized measures of reading comprehension to informal assessments of reading skills including ORF, maze reading, and teacher judgment. The subjects in this study included 335 students in grades 2 - 6. According to the authors, correlations between ORF and performance on the Gates-MacGinitie Reading Tests and Metropolitan Achievement Test (MAT) declined in successive grades with correlations of .83 - .86 in 2nd grade to correlations of .58 - .67 in 6th grade. In grades 3 - 5, correlations remained strong with most above .7.

In 2005, Hosp and Fuchs analyzed the relationship between ORF passages and standardized assessments of reading skills. Participants in the study included 310 students in grades 1 -

4 from an urban school district assessed to determine if the relationship between CBM measures and specific reading skills were impacted by the grade level of the students assessed. Two CBM reading passages, developed by Fuchs and Fuchs (as cited in Hosp & Fuchs, 2005), were administered to students at each grade level. The average number of wcpm for the two passages was used as the representative CBM score. Participants were also assessed with the Work Attack, Word Identification, and Passage Comprehension subtests on the Woodcock Reading Mastery Test – Revised (WRMT-R) The Basic Skills and Total Reading (Short version) indices on the WRMT-R were calculated using the subtests assessed. According to the authors, results of their study indicated strong relationships were present between ORF with standardized measures of decoding, word identification, basic reading skills, reading comprehension, and an overall reading composite at each grade level measured. Correlations between ORF and the three WRMT-R subtests were strong. The correlations between ORF and decoding ranged from .71 - .82. The correlation between ORF and Word Reading ranged from.73 - .91 with correlations ranging from .79 - .84 between ORF and Reading Comprehension. Similar correlations were found between ORF and the two WRMT-R indices, Basic Skills and Total Reading (Short version). On

these measures, correlations ranged from .78 - .89 and .83 - .91 respectively.

Hit rates, sensitivity and specificity were also calculated to determine success on each of the reading skills and two indices (Hosp & Fuchs, 2005). The authors suggest that ORF was able to accurately distinguish students who fell above or below standard scores of 85 – 90 on each of the WRMT-R measures. Cut scores were also generated for ORF at each grade level using standard scores on each of the WRMT-R subtests and both indices as the criteria for successful reading. The authors recommend using cut scores for ORF based on the Basic Skills and Total Reading indices not the individual subtests. This is because cut scores derived from the indices provided a better estimate of overall reading achievement.

Wayman, Wallace, Wiley, Ticha, and Espin (2007) conducted a literature synthesis on the use of curriculum-based measures of reading. They selected 160 research studies that met their predetermined criteria. They identified all of the articles addressing CBM in reading, writing and math in kindergarten through grade 12 using several electronic databases. This initial search produced 160 documents, 90 of which referred for CBM in reading. From these 90 articles they examined only studies that were completed after 1989 and that addressed

issues of the technical adequacy of CBM in reading. Finally, they chose studies which evaluated three specific curriculum-based reading measures including reading aloud, maze selection, and word identification. This final selection narrowed the field down to 65 studies.

Wayman et al. (2007) concluded that CBM ORF measures are a valid procedure when used by educators as a measure of general performance. Many of the research studies evaluated highlighted positive relationships between ORF and a variety of measures of reading comprehension across various settings and situations. The authors suggest that the use of ORF as a measure of general reading outcomes is supported by their research findings.

Shinn et al. (1992) evaluated the relationship between ORF measures and the general reading process for students in grades 3 and 5 using confirmatory factor analysis. The purpose of their study was to show that the concept of ORF fit into contemporary theoretical models of reading development. Additionally, they sought to show that ORF is a reliable and valid measure of general reading comprehension. The authors recruited 114 third-grade and 124 fifth-grade students from a public school district in a mid-size northwestern city. The sample was predominately white with 96% of students receiving instruction within the general education setting. Each

subject was assessed with eight separate measures of reading. Two of the measures assessed reading decoding, the Test of Written Spelling (TWS) and Word Attack subtest from the WRMT. Four reading comprehension measures were administered including the Literal and Inferential Reading Comprehension subtests from the SDRT, a cloze reading task, and a written retell task. Finally subjects were assessed with two ORF passages developed from the Harcourt-Brace Jovanovich basal reading series.

To complete the study, Shinn et al. (1992) compared the fit of four models to the collected data. First, a unitary model was evaluated where fluency, decoding, and comprehension were not distinct. Second, the authors evaluated a two-factor model of decoding and comprehension that included reading fluency as part of the construct of decoding. Third, a two-factor model of decoding and comprehension was analyzed that included reading fluency as part of the reading comprehension construct. Finally, a three-factor model of reading where reading fluency was a separate construct was evaluated.

According to Shinn et al. (1992) a unitary model of reading was validated with significant contributions from all measures, including ORF, into the model in the third-grade sample. This model hypothesized that each measured variable represent the latent variable Reading Competence. Factor

loadings on the Reading Competence construct ranged from .68 for written retell and .90 for ORF. Goodness of fit indices indicated an adequate fit to the data for the one-factor model. Both two- and three-factor models did not offer significant improvement in fit.

A two-factor model of reading that included fluency representing decoding was validated for fifth-grade students (Shinn et al., 1992). Loadings on Reading Comprehension in this model ranged from .61 for written retell to .86 for cloze measures. Factor loadings on the Reading Decoding construct ranged from .66 for nonsense words to .90 for ORF. Goodness of fit indices was greater than .90 for the two-factor, fluency as decoding model. No other model provided an improvement in goodness of fit.

Shinn et al. (1992) concluded that their findings are consistent with contemporary theories of reading assessment, particularly CBM. Of particular importance, they found that no matter what factor model was employed, ORF was confirmed as a valid measure of reading comprehension skills. Although debate regarding the use of ORF as a measure of reading comprehension continues, primarily due to face-validity concerns, they suggest that these arguments should be put to rest as ORF has shown to be a valid measure of reading comprehension. They recommend efforts to improve the already

strong validity of CBM to increase its utility as a problem-solving tool.

Each of these studies included in this section show moderate to strong correlations between ORF and overall reading achievement with most coefficients near .70. These correlations were strong regardless of the standardized achievement test used to measure overall reading achievement. Additionally, evidence was discussed showing that the correlation between ORF and reading achievement decreases in the upper elementary grades. Correlations remain strong, however, in grades 3 – 5 included in this study.

The following section further analyzes and extends the relationship between CBM and reading achievement. Included are several studies which examine the correlations between CBM and high stakes assessments. In addition, the calculation and use of locally-generated ORF benchmarks are discussed as well as a determination of the accuracy of these benchmarks for predicting proficiency on the state test.

CBM and high-stakes testing. Given its research supported relationship with general reading outcomes, CBM has been routinely utilized to predict performance on high stakes assessments (Deno, 1985; Silberglitt, 2008). As the pressure for school districts to improve test scores increases, teachers and administrators have used formative CBM to adjust

student curricula, instructional practices, and supplemental strategies as a means to help students meet proficiency on the state assessment. Over the past decade, a proliferation of research studies measuring the link between CBM performance and performance on state assessments has been explored.

Crawford, Tindal, and Steiber (2001) evaluated the utility of curriculum-based measures, particularly ORF, to predict performance on statewide achievement tests. Fifty-one students in grades 2 and 3 participated in the study. The students were primarily white with 29 girls and 22 boys. Only 9 of the students received special education services. The researchers in this study created reading passages taken from the Houghton Mifflin Reading Series. Students were assessed in grades 2 and 3 using the developed reading probes. The scores earned on these measures were compared to the results they earned on the Oregon Statewide Reading Assessment at the end of third grade. The results of their analysis showed moderate correlations were identified between the Houghton Mifflin reading probe scores and performance on the Oregon Statewide Assessment (.60 in grade 3 and .64 in grade 2). Further, when students earned an ORF score above a pre-established benchmark they were extremely likely to meet proficiency on the Oregon Statewide Assessment. The authors suggest that the use of ORF to predict performance on state

assessments and the use of CBM to monitor student's growth toward benchmark goals is supported.

As part of the validation study for DIBELS 6th Edition, Shaw and Shaw (2002) analyzed the relationship between performance on DIBELS ORF and scores on the third-grade Colorado State Assessment Program (CSAP). They assessed 52 third-grade students from a Colorado elementary school. The study took place during the fall, winter, and spring of the 2001 – 2002 school year. The authors reported descriptive statistics for DIBELS ORF and the CSAP at each time period during the school year as well as correlations between ORF and the CSAP. Correlations between the DIBELS ORF and performance on the CSAP were .73 in the fall, .73 in the winter, and .80 in the spring. Additionally, they found that 91% of third-grade students who took part in this study scored proficient or advanced on the CASP when they were able to read at or above 90 wcpm on DIBELS ORF. They also found that 73% of students who earned a DIBELS ORF score below 90 failed to meet proficiency on the CSAP. Ninety percent of students who scores above the 110 wcpm met proficiency on the CSAP with 43% of students falling below 110 failing to meet proficiency on the CSAP. This analysis suggests that the benchmark scores generated by the DIBELS system can be used to predict performance on a state assessment.

Good, Simmons, and Kame'enui (2001) also measured the validity of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) assessment to predict performance on a state assessment, the Oregon Statewide Assessment. In this study, four cohorts of students in grades 3 - 6 were selected from an urban school district in the Pacific Northwest. There were approximately 350 students in each cohort. Ten percent of the participants were considered a racial or ethnic minority. Eighteen percent of students were considered living at or below the poverty line. They found correlations as high as .82 and were able to show that 96% of students who met the ORF benchmark goal in grade 3 were able to meet proficiency on the Oregon Statewide Assessment. According to the authors, their findings support the use of benchmark goals to evaluate and identify students at-risk for failure and adjust their instruction so they are able to attain reading proficiency.

The connection between the DIBELS and performances on the Iowa Tests of Basic Skills (ITBS) was evaluated by Schilling, Carlisle, Scott, and Ji (2005) to determine whether fluency measures are accurate predictors for reading achievement. Data were collected in nine school districts that made up the first Reading First cohort. The Reading First initiative was enacted to provide low-achieving and high-poverty school with resources to help students become proficient readers.

Approximately, 2,500 students participated at each grade level. The majority of students identified as African American (60%) with 24% identified as white, and 13% identified as Hispanic. Eighty-one percent of students were determined to be economically disadvantaged. Fifteen percent were described as Limited English Proficient and 8.5% of students received special education services.

Students were assessed in the fall winter and spring with the following DIBELS indicators: Letter Naming Fluency, Nonsense Word Fluency, Word Usage Fluency, Phoneme Segmentation Fluency, and ORF (Schilling et al., 2005). In addition, students were assessed with the ITBS which measures vocabulary, word analysis, listening comprehension, language (grammar and spelling), and reading comprehension. The authors found correlations between ORF and the ITBS subtests were moderate to strong ranging from .61 - .75 in all areas except Listening (.31). These correlations outperformed those obtained for the ITBS subtests and the other DIBELS indicators assessed with all correlations below .6. The validity of the DIBELS Benchmark scores to predict successful performance (defined as the 50th percentile) on the ITBS was also analyzed in this study. Hierarchical regression was used to determine how well DIBELS indicators at each testing time predicted Spring ITBS scores. Their findings show that the DIBELS

significantly predicted performance on the ITBS. They found that 80% of second-grade students and 76% of third-grade students who fell in the *At-Risk* category in ORF also fell in below the 25th percentile on the ITBS. However, they also found that the *Low-Risk* category made for a less accurate prediction of proficiency with only 32% of second-grade students and 37% of third-grade students who were identified as *Low-Risk* on DIBELS did not meet proficiency (50th percentile) on the ITBS. This is problematic because educators place trust in the DIBELS benchmark to accurately distinguish the students who need additional support from those who do not. False negatives, like those found by Schilling et al. (2005), can be harmful as students may be denied important services after receiving an incorrect identification of *Low-Risk*. To correct this concern, the authors recommend more frequent progress monitoring with a variety of assessments to measure student progress. This will help to ensure accurate decisions are made regarding the services students receive.

In 2001, Stage and Jacobson evaluated the relationship between ORF assessed in September, January, and May, and performance on the Washington Assessment of Student Learning (WASL) assessed at the end of the school year. One hundred seventy-three fourth-grade students from a one elementary

school in Puget Sound participated in the study. Fifty-four percent of the students were male. Ninety percent of the total number of students identified as European America. Fifteen percent of the population was eligible for free and reduced lunch. Eleven of the 174 participants received special education services.

Stage and Jacobson (2001) developed reading probes using the students' current reading curriculum. After administering the probes to students, the researchers were able to compare the CBM scores to the scores earned on the WASL. Cut scores were developed using three analyses of variance (ANOVA) with proficiency on the WASL as the criterion for successful achievement. Stage and Jacobson used ANOVA to calculate estimated mean ORF scores along with a 95% confidence interval for three time periods during the school year: fall, winter and spring. The score at the lower end of the 95% confidence interval during each of these time periods was selected as the cut score.

Diagnostic efficiency statistics were calculated for the cut scores with pass or fail on the WASL as the criterion (Stage & Jacobson, 2001). Diagnostic accuracy statistics were produced for the September cut score. This analysis yielded sensitivity of 66% with specificity of 76%. PPP was 41% with NPP of 90%. These statistics were also calculated for the

January and may cut scores. There was less than a one percent difference between the diagnostic efficiency statistics of three cut scores indicating nearly equal results. Overall accuracy of the cut scores was determined using gamma and kappa. Analysis with gamma yielded an overall accuracy percentage of 74%. Analysis with kappa, which corrects for chance agreements, yielded diagnostic efficiency 34% above chance.

According to Stage and Jacobson (2001), there is a great deal of potential for using ORF cut scores to predict performance on a state-mandated test. By using and ORF cut score, educators can identify which students are at-risk for reading failure and adjust their instruction to meet these students' needs. Students who are making inadequate progress in relation to the calculated cut scores can be referred

McGlinchey and Hixon (2004) conducted a replication of the work of Stage and Jacobson (2001). They evaluated the predictive value of CBM performance on the Michigan Educational Assessment Program (MEAP) for fourth-grade students. The study took place primarily in one elementary school in an urban school district over an eight year period. During year four, the entire school district's fourth grade consisting of 14 elementary school building participated in the study. The elementary building that participated for all

eight years had between 55 – 139 students enrolled per year. A total of 1,326 students participated during the entirety of the study within this elementary school building.

ORF measures were developed from passages selected randomly from the Macmillan Connections Reading Program (McGlinchey & Hixon, 2004). Only one 1-minute reading probe was administered to the participants in the first five years of the study. Three 1-minute reading probes were assessed in the remaining three years due to the availability of increased staff support. When three passages were assessed, the median score was selected to represent the ORF score. The ORF measures were assessed two weeks prior to the completion of the MEAP.

To calculate diagnostic efficiency statistics, McGlinchey and Hixon (2004) selected a score of 100 wcpm as the cut score for proficiency. This cut score was selected because of previous research findings which identified 100 wcpm as an appropriate cut score for this grade level. The specificity of the cut score for identifying students who met proficiency was 74%. The sensitivity of the cut score for students who did not meet proficiency was 75%. PPP, which identified the probability of students who failed the MEAP, was 77%. The NPP of the cut scores, which represented the probability of accurately identifying students who passed the MEAP, was 72%.

The overall classification accuracy was 74%. Kappa was also calculated to measure diagnostic efficiency. This analysis showed that the efficiency for the cut score was 48% above chance. In addition, the authors found that ORF is highly predictive of performance on the MEAP with correlations ranging from .63 - .81. They suggest targets can be developed using this relationship between CBM and state assessment performance to identify which students are on track and which are not. Once identified, teaches can adjust the instruction to meet struggling students' needs.

Buck and Torgesen (2003) evaluated student scores in ORF to performance on the Florida Comprehensive Assessment Test – Sunshine State Standards (FCAT-SSS). Data were collected from 13 schools in one Florida school district and included 1102 students. The FCAT-SSS was administered in the April of 2002 and the students were assessed with ORF in May of 2002. The Standard Reading Passages: Measures for Screening and Progress Monitoring from Children's Educational Services were used to measure ORF. Their analysis showed that ORF can very accurately predict scores on the FCAT-SSS for a heterogeneous group of third-graders. A correlation of .70 between ORF and reading the FCAT-SSS was identified. This finding is consistent with previous research. Buck and Torgeson conducted a similar analysis as was done by Shaw and Shaw

(2002) with comparable results. They found that 91% of students who were able to read above 110 wcpm were able to perform successfully on the FCAT-SSS. Conversely, 81% of third-grade students who read below 80 wcpm performed unsatisfactorily on the FCAT-SSS.

Hintze and Silberglitt (2005) evaluated the relationship between CBM of reading and performance on the Minnesota Comprehensive Assessment (MCA) for students in grades 1 - 3. Participants included 1,766 students from seven school districts in the north central United States. Fifty-one percent of the participants were males and the vast majority of the participants were identified as White not of Hispanic origin. Approximately 5% of the students received special education services.

Students in grades 1 – 3 were assessed with the AIMSweb R-CBM passages in the fall, winter, and spring over a three year period (Hintze & Silbergiltt, 2005). In addition, students were assessed using the reading portion of the MCA in the spring of third grade. The authors developed R-CBM cut scores using end of the year third-grade MCA as the outcome measure for all CBM assessments. Three different methods were used to calculate the cut scores including logistic regression, discriminant analysis, and receiver operator characteristic (ROC) curve analysis. Additionally, the

evaluated whether using the MCA as a constant predictor for each cut score or whether using successive R-CBM benchmarks as a predictor of proficiency produced better results.

Hintze and Silberglitt (2005) found that R-CBM has a strong relationship with performance on the MCA as much as two years in advance. Correlations between the MCA and R-CBM in assessed in third grade ranged from .66 - .69. Similar correlations were found between the MCA and second-grade R-CBM (.61 - .68). Correlations of .49 and .58 were identified between first-grade R-CBM and the MCA.

Evaluation of the ROC curve analysis, discriminant analysis, and logistic regression produced consistent results (Hintze and Silberglitt, 2005). All three methods generally yielded higher levels of specificity and PPP as opposed to sensitivity and NPP. For example, the following cut scores and diagnostic accuracy statistics were reported for each of three methods used for calculating cut scores. Theses cut scores were developed using successive benchmarks as the criterion for proficiency. ROC curve analysis produced a cut score of 76 in the spring of third grade (Sensitivity = .88, Specificity = .94, PPP = .62, NPP = .99). A cut score of 112 was created using discriminant analysis (Sensitivity = .95, Specificity = .77, PPP = .87, NPP = .93). Logistic regression yielded a cut score of 81 (Sensitivity = .85, Specificity =

.93, PPP = .84, NPP = .93). The authors suggest that these comparable levels of diagnostic efficiency indicate that R-CBM is a strong predictor of MCA performance regardless of the method of cut score calculation.

Hintze and Silbergiltt (2005) state that CBM is an efficient procedure for predicting performance on state tests as accurate predictions are able to be made as much as two years in advance. Their results also show that each of the three procedures used to create cut scores show acceptable levels of specificity and sensitivity. The use of any one will depend on the potential use of the cut score within the school district. Based on the balance between sensitivity and specificity, the authors suggest that one method may produce a score which is best used for screening while another would be better used in special education decision making. The authors also reported that setting the cut scores using a successive method from one benchmark to the next produced more accurate and efficient cut scores than when using the MCA as the only criterion. They suggested that this is "because predictions made from one benchmark period to the next occur more closely in time to each other than to the MCA, which, depending on the benchmark in question, can be far removed in time" (p. 383).

Silberglitt, Burns, Madyun, and Lail (2006) conducted a longitudinal analysis of the relationship between reading

fluency data and state accountability test scores. The purpose of their study was to examine whether the relationship between CBM and state test performance was a function of the grade level assessed. Data for 5,472 students from five school districts in rural and suburban Minnesota were collected over a period of 7 years. Fifty-one percent of the participants were male and the vast majorities (94.3%) were identified as white, not of Hispanic origin. Approximately 5 – 18% of the students in the five school districts lived in households that met the federal definition for poverty. Each student in grades 3, 5, 7, and 8 were given the Minnesota Comprehensive Assessment – Reading (MCA-R), the Minnesota state accountability assessment, in the spring of each year. Students were also administered the AIMSweb R-CBM, AIMSweb Maze, and the Basic Standards Test-Reading (BST-R).

Silberglitt et al. (2006) calculated correlations between the three formative assessments (R-CBM, Maze, and BST-R). They found that the relationship between ORF and performance on the Minnesota Comprehensive Assessments - Reading (MCA-R) diminished as the grade level advanced. This pattern was exemplified as the correlation between CBM-R and the MCA-R showed a steady decline from .68 in grade 3 to .60 in grade 7. The authors suggested that educators should use caution when using ORF measures to predict state test performance,

particularly in the later grades. Similarly establishing target scores based on success on a state test may be inappropriate in the later grades. The authors recommended using other assessments tools other than ORF data to monitor progress toward the end of the year state test.

The predictive validity of ORF, Maze, and the combination of both measures to the Measures of Academic Progress (MAP) in Nebraska was analyzed Beckman and Merino (2010). The authors collected data from 376 students in grades 2 – 5 from one elementary school in Nebraska. The participants in the study were from various backgrounds including English Language Learners (39%), students with disabilities (15%), Caucasian (20%), African (4%), Hispanic (64%) and others (2%). Each student was assessed with the AIMSweb R-CBM and Maze tests in the fall and spring of the school year. In addition the participants were also assessed in the fall and spring with the MAP.

Correlations of .62 (grade 3), .66 (grade 4), and .66 (grade 5) were identified between spring AIMSweb R-CBM and the fall MAP of the following school year (Merino & Beckman, 2000). The authors found that R-CBM alone and the combination of ORF and Maze significantly predicted performance on the MAP in grades 2 - 5 ($p < .05$). Maze without the addition of R-CBM did not significantly predict MAP performance in all grade

levels. According to the authors, the results support the use of ORF measures like AIMSweb R-CBM to monitor student who are at-risk for failing their state assessments. They suggested that ORF is a better predictor of state test performance than maze in grades 2 - 5. The combination of ORF and maze, however, does create a stronger prediction than either measure by itself.

The studies included in this section show that curriculum-based measures of ORF show a strong connection with proficiency on a state assessment. This relationship is strong regardless of the state assessment utilized or the population assessed. As with the findings between ORF and standardized achievement tests in reading, correlations between ORF and state assessments are strong to moderate with coefficients ranging from .61 - .69. In addition, studies that employed the use of CBM benchmarks developed using state test performance as a criterion for successful reading outcomes demonstrated that these cut scores can be used as an accurate predictor of state test performance. In the following section, the relationship between CBM and state test performance is further evaluated to determine if similar findings will be shown using the Pennsylvania System of School Assessment (PSSA).

CBM and the PSSA. Shapiro, Keller, Lutz Santoro, and Hintze (2006) evaluated the relationships between CBM of reading and math to the PSSA along with three additional standardized achievement tests. Data were collected from second-grade and fourth-grade students from two school districts in eastern Pennsylvania during the 2002 – 2003 school year. Participants were selected using stratified random sampling based on the socioeconomic status of the students. This was done so that the sample would reflect the economic background of the school district population (32% free and reduced lunch). This process yielded 617 students from District 1 with 782 students from District 2. The participants were administered the AIMSweb R-CBM in the fall, winter, and spring of the school year according to the standardized directions. Students were also assessed with the PSSA along with three other standardized achievement tests: the Stanford Achievement Test – Ninth Edition (SAT-9), the Metropolitan Achievement Test – Eighth Edition (MAT-8) and the Stanford Diagnostic Reading Test (SDRT). Each of the standardized achievement tests were assessed in the spring of the school year. Correlations were calculated Between R-CBM and each of the standardized achievement tests. The results indicated moderate to strong correlations (.70 - .74) between fall, winter, and spring R-CBM and both the SAT-9 and the MAT-

8 assessed at the end of the year. Correlations between R-CBM assessed in the fall, winter, and spring and the SDRT were not as strong with correlations of .50, .52, and .55. The relationship between R-CBM assessed in the fall, winter, and spring and the PSSA were also moderate to strong with correlations of .65 - .69. According to the authors, their findings suggested that CBM data may be appropriately used to help determine which students will be successful on their end of the year state assessment.

In 2008, Keller-Margulis, Shapiro, and Hintze examined the long-term diagnostic accuracy of CBM predictions of PSSA proficiency. They analyzed student CBM scores in reading and math and predicted their performance on the PSSA using locally-generated cuts cores 1 - 2 years prior to taking the test. They selected 1,461 students from one school district in eastern Pennsylvania. The Population of the school district included students from different racial/ethnic groups including Caucasian (58%), Hispanic (31%), African American (9%), and Asian (3%). Approximately 33% of the student population resides in a low-income household as determined by free and reduced lunch status.

According to the authors, correlations were significant between second-grade ORF scores and PSSA scores at the end third grade (.97 - .71) as well as between ORF scores in grade

4 and PSSA scores at the end of grade 5 (.67 - .69) (Keller-Margulis, Shapiro, & Hintze, 2008). Significant correlations were also found between ORF scores and PSSA scores assessed two years later with correlations of .53 - .69 between first-grade ORF and end of the year third-grade PSSA scores along with correlations of .72 - .74 and between third-grade ORF to end of the year fifth-grade PSSA scores. The authors noted that they were able to accurately predict students' PSSA performance levels as pass or fail as much as 84% of the time using CBM. They suggested that their data highlights the importance of screening to ensure that students receive supports as early as possible to intervene with developing academic concerns.

In 2008, Shapiro, Solari, and Petscher evaluated the diagnostic accuracy of predictions made by DIBELS ORF and the 4Sight Benchmark Assessment (Success for All Foundation, 2007) for proficiency on the PSSA in grades 3 - 5. The 4Sight is a 30-item group-administered test. The reading portion of the 4Sight was based on the Pennsylvania curriculum standards and assessment anchors with a particular focus on reading comprehension. It was designed to be predictive of proficiency on the PSSA. The authors also combined DIBELS ORF with performance on the 4Sight to determine if the combination would improve the accuracy of the prediction. According to

the authors, correlations ranging from .64 - .75 were found between ORF and the PSSA across grades 3 - 5 with correlations of .71 - .75 between the 4Sight and the PSSA. Both DIBELS ORF and the 4Sight assessment demonstrated good levels of prediction with the PSSA by themselves. ORF with the combination of 4Sight and DIBELS ORF scores was able to outperform the predictions of PSSA performance made by either assessment alone.

Goffreda, DiPerna, and Pedersen (2009) assessed the predictive validity of the DIBELS indicators in assessed in first grade to PSSA proficiency in third grade. Using logistic regression, they sought to predict student scores on the TerraNova California Achievement Test (CAT) and the PSSA. They assessed each of the DIBELS indicators assessed in middle of the year in first grade including Nonsense Word Fluency (NWF), Phoneme Segmentation Fluency (PSF), Letter Naming Fluency (LNF), and ORF. DIBELS scores were compared to scores earned on the CAT at the end of second grade and the PSSA at the end of third grade. Sixty-seven students from one school district in central Pennsylvania participated in this study. Fifty-six percent of the participants were male, with 78% White, 10% Hispanic, 2% Black, and 1% Asian according to the authors.

Correlations between the LNF, PSF, and NWF first-grade DIBELS measures and the CAT and PSSA were moderate (Goffreda et al., 2009). The relationship between ORF and the CAT and PSSA were also moderate with correlations of .39 and .54, respectively. Logistic regression was conducted to determine if the DIBELS risk categories (*At-Risk, Some-Risk, Low-Risk*) significantly predicted PSSA and CAT proficiency. When all 4 DIBELS measures were combined they significantly predicted proficiency on the CAT and PSSA. When the DIBELS measures were analyzed separately, only ORF significantly predicted proficiency on the CAT and PSSA.

Cut scores using ROC curves were generated by Goffreda et al., (2009) which maximized sensitivity and specificity. These cut scores were compared to those recommended by the DIBELS system. None of the DIBELS provided cut scores for the four DIBELS measures provided adequate (>80%) levels of sensitivity and specificity. The cut scores generated by the authors for LNF, PSF, and NWF also did no show adequate levels of sensitivity and specificity. The ORF cut scores created by the authors showed adequate levels of sensitivity and specificity for both the CAT (80% and 87%) and the PSSA (88% and 88%).

In their poster presentation, Ferchalk et al. (2010) assessed the predictive validity of DIBELS ORF for performance

on the PSSA in third grade. They evaluated three years of archival DIBELS benchmark data gathered for 576 third-grade students in a rural school district in Pennsylvania. Student benchmark scores in the fall, winter, and spring were compared to the PSSA scores earned at the end of the year. Once gathered, the data were entered into a variation on a slope intercept equation X = (Y-a)/b where Y = proficiency on the PSSA, a = the Intercept, b = slope.

The benchmarks that were developed using this procedure were applied to a new cohort of students and compared to the benchmarks generated by the DIBELS system (Ferchalk et al., 2010). The overall correct percentage for the fall, winter, and spring locally-generated benchmarks was an average of 83%. The DIBELS benchmarks accurately predicted pass/fail on the PSSA an average of 78% of the time. The locally-generated benchmarks produced low levels of sensitivity (.60) with very high levels of specificity (.96). The DIBELS benchmarks produced an opposite trend with higher levels of sensitivity (.89) with lower levels of specificity (.72).

Ferchalk, Cogan-Ferchalk, and Richardson (2012) assessed the correlations between the 4 DIBELS indicators used in third grade including Daze, ORF, Retell Fluency, and Reading Accuracy and the PSSA. Additionally, correlations were calculated between the recently developed DIBELS composite

score, which includes 4 DIBELS indicators, and the PSSA. The authors analyzed data for 184 students across 4 elementary school buildings in one school district. Each participant was assessed with the DIBELS ORF in the fall, winter, and spring and was also assessed with the PSSA in the spring. Correlations were calculated between each of the DIBELS indicators at each of three time periods with the PSSA. ORF produced the strongest correlations (.68 - .70) with the PSSA followed by Daze (.60 - .70), and Accuracy percentage (.63 - .68). Correlations between the retell fluency measures (Retell Fluency and Retell Quality) and the PSSA were weaker with correlations ranging from (.52 - .61). The DIBELS composite score outperformed ORF and the individual DIBELS indicators with correlations ranging from .75 - .78. Given the available research on the Retell Fluency and the Daze assessments, the authors developed several additional composite scores, which reduced the weight of these indicators within the composites. Contrary to the authors' hypothesis, none of the researcher-generated composite scores demonstrated stronger correlations with the PSSA than the DIBELS composite score. Pending further research, the authors suggested that the DIBELS composite score may be a valid indicator of performance on the PSSA.

In this section, the relationship between ORF and performance on the PSSA was evaluated. Correlations between the ORF and PSSA were moderate to strong with most studies reporting coefficients near .70. This finding is consistent with the research studies connecting CBM with other measures of reading achievement including other state assessments and standardized measures of reading achievement. Three studies include in this review also evaluated the use of locally-generated benchmark scores for predicting state assessment performance. All three studies found that these locally-generated benchmarks appear to be valid targets for use in predicting performance on a state assessment.

In all three of the previous sections strong correlations were found between CBM and performances on standardized achievement tests, various state assessments and the PSSA. It is important to note that perfect correlations were not present between CBM and any assessment presented in this review. In many of the studies reviewed, correlations near .7 were found. Although strong, this suggests that 49% of the variance was explained by the CBM leaving a noticeable amount of variance unaccounted. According to Hasbrouck and Tindal (2006) some variance can be expected due to extraneous factors including student interest in the assessment activity, student background knowledge, accuracy of assessment, assessment

errors. In addition difference in the educational environment including teacher characteristics and instructional practices may also impact the relationship between CBM and overall reading proficiency. These extraneous factors must be considered when interpreting the relationship between CBM and overall reading achievement as well as any extension of this relationship including benchmark scores.

Even after considering the extraneous factors that may interfere with the interpretation of CBM, The research presented in this review provides strong evidence for the development and use of locally-generated ORF benchmark scores to predict performance on a state assessment. This is substantiated by the strong correlation between ORF and reading achievement as well as previous successful predictions made between local ORF benchmarks and state test performance. The present study applies these relationships to create a set of locally-generated benchmarks using the DIBELS and the PSSA. In the following sections both measures used in this study are presented in greater detail with a discussion of their characteristics and uses as well as their reliability and validity.

Dynamic Indicators of Basic Early Literacy Skills

The DIBELS are a series of fluency based assessments designed to efficiently monitor student progress throughout

the school year toward end of the year goals (Good et al., 2011b). Students are assessed three times a year in the fall, winter and spring. The DIBELS are currently on their 7th edition titled DIBELS Next. During each assessment students are asked to demonstrate their reading skill on one or more indicators for a short period of time, 1 - 3 minutes depending on the indicator. The scores earned by students on the DIBELS are then compared to benchmark expectations, or cut scores, which indicate the likelihood that a student will reach subsequent reading goals.

The DIBELS assessments are comprised of several individual tests including Phoneme Segmentation Fluency, Letter Naming Fluency, First Sound Fluency, Nonsense Fluency, ORF, and Daze (Good, Kaminski, Dewey et al., 2011). In grades 3 – 5, only ORF, with its supplemental and optional retell fluency and reading accuracy components, and Daze, a maze reading procedure, are assessed.

DIBELS Oral Reading Fluency. DIBELS ORF is designed to assess a student's reading fluency which is defined as the effortless, automatic ability to read words in connected text (Stahl, 2004; Torgesen & Hudson, 2006). Fluency is important to emergent reading skills because when children are able to read words efficiently their reading ability is no longer an obstacle to their ability to gain meaning from what they read

(Torgesen & Hudson, 2006). During universal screening with DIBELS ORF, students are asked to read aloud from 3 separate reading passages for 1 minute each (Good et al., 2011b). The median number of words correctly read per minute (wcpm) on the three passages is used as the primary metric to determine student performance (Shinn, 2008).

Two supplementary metrics, passage Retell Fluency and reading accuracy, are optional indicators incorporated into DIBELS ORF passages. Passage Retell Fluency is designed to be a brief check of comprehension and is assessed following each 1 minute ORF passage (Good et al., 2011b). After the ORF passage is completed, the evaluator asks the examinee to retell what just read. The number of words relevant to the story the student was able to retell are recorded. Retell Fluency includes an additional indicator that instructs evaluators to make a judgment from 1 - 4 regarding the quality of the retell fluency response. The second supplementary metric, accuracy percentage, provides an indication of how accurate or how many errors the student made when reading the passage. The accuracy percentage for DIBELS ORF is calculated by dividing the median number of words that the student attempted to read by the median number of words the student correctly read. Both of these additional metrics provide further information regarding student performance on DIBELS

ORF. In addition, their inclusion increases the reliability and validity of ORF. Retell fluency and reading accuracy ensure that students are reading the ORF passages for meaning, not just speed (Good et al., 2011b)

Daze. Daze is a maze procedure that has been standardized by the DIBELS system to assess reading comprehension (Good et al., 2011a). A maze procedure assesses the reasoning skills that accompany reading comprehension and a student's ability to form meaning from what they have read (Shapiro et al., 2008). In the format created by the DIBELS system, approximately every 7th word in a reading passage is eliminated and replaced with a three-item multiple choice selection (Good et al., 2011a). Students are then asked to read the passage for 3 minutes and select the choices that best fit into the sentences. The number of correct words minus half the number of incorrect words is recorded as the student's Daze adjusted score. This score is used as an estimate of a student's level of reading comprehension.

DIBELS composite score. The DIBELS composite score is an amalgamation of several separate indicators and, according to the authors, provides a more reliable measure of students' reading skills than any of the individual indicator alone (Good et al., 2011a). In grades 3 – 5, the DIBELS composite score is comprised of four different indicators. Three of the

four indicators are derived from the ORF measure including the number of words students read correctly per minute (wcpm), the number of words that a student correctly retells after reading the passage (Retell Fluency), and the student's percentage of words read correctly (Accuracy Percentage). In addition, the student's adjusted score on the DAZE is added into the composite. The scores are combined to create the DIBELS composite score using the following calculation: ORF wcpm + (RTF x 2) + (Daze x 4) + Converted Accuracy percentage. An accuracy percentage conversion table is available in the DIBELS benchmark goals and composite score supplement (Dynamic Measurement Group, 2010). Based on their research, the authors of DIBELS found that the composite is more predictive of general reading outcomes than any of the individual indicators including ORF.

Reliability and validity of DIBELS. Validation studies of DIBELS Next were conducted by the Dynamic Measurement Group (DMG), the agency who produces the DIBELS measures (Powell-Smith, Good, Latimer, Dewey, & Kaminski, 2011). The DMG analyzed test-retest reliabilities of ORF wcpm in grades 3 - 5. They selected 120 predominately white (94%) students from one school district in the northwest region of the United States. The students were assessed with the mid-year

benchmark and again two weeks later. Test-retest reliability ranged from .93 - .97.

Inter-rater reliability for DIBELS ORF was evaluated using a sample of 122 predominately white (94%) students across five schools in the northwest and Pacific west of the United States. To evaluate inter-rater reliability, a shadow-scoring procedure was used. Several evaluators, after receiving training in the scoring procedures, scored the ORF of the same student. The results indicated that inter-rater reliability of ORF was high at .99 for all grade levels assessed.

Powell-Smith et al. (2011) established DIBELS criterion-related validly by determining the extent to which the student performance on the Group Reading Assessment and Diagnostic Evaluation (GRADE) was predicted by DIBELS scores. The authors selected 3,190 students from five predominately white (94%) school districts in the north central Midwest and the Pacific west United States. Predictive validity was calculated by comparing end of the year GRADE results with scores earned on DIBELS in the fall and winter. The results of this comparison yielded coefficients ranging from .66 - .77 in grades 3 - 5. Concurrent validity was calculated by comparing scores on the GRADE assessed in the spring to the

DIBELS ORF score also assessed in the spring. Concurrent validity coefficients for ORF ranged from .65 - .74.

Pennsylvania System of School Assessment

The purpose of the PSSA is to provide educators, students, parents, and members of the community with knowledge about student and school performance according to the Data Recognition Corporation (DRC; 2011), the company that produces the PSSA. It is a standards-based, criterion-referenced measure designed to determine the extent school curricula help students to attain proficiency of academic skills. The reading portion of the PSSA measures five general academic skills including learning to read independently; reading critically; reading, analyzing, and interpreting fiction; characteristics and functions of the English language and research (Shapiro et al., 2006). To assess these skill areas, students are asked to independently read a series of passages and answer questions which correspond to one or more of these five general skill areas.

The PSSA produces a scaled score which ranges from 700 to 2100 (Shapiro et al., 2006). Proficiency cut points for each grade level measured in this study are as follows: Grade 3 = 1235, Grade 4 = 1255, Grade 5 = 1275 (DRC, 2011). To validate the cut scores, a modified version of a bookmarking procedure was used (DRC, 2011). A panel of educators in Pennsylvania

was presented with a booklet containing test items from the 2005 PSSA assessment. The items were arranged from easiest to the hardest determined by number of students who answered each item correctly. The panel was asked to bookmark the items where the borderlines between each of the performance level descriptors would fall. Test items that precede the placed bookmark indicate skills which all students within that performance level should know. The panel completed this procedure several times with ongoing discussions to ensure consensus between members and the validity of the cut scores. The following performance level descriptors are also assigned to categorize student performance:

> Advanced. The Advanced Level reflects superior academic performance. Advanced work indicates an in-depth understanding and exemplary display of the skills included in the Pennsylvania Academic Content Standards (DRC, 2011, p. 218).
>
> Proficient. The Proficient Level reflects satisfactory academic performance. Proficient work indicates a solid understanding and adequate display of the skills included in the Pennsylvania Academic Content Standards (DRC, 2011, p. 218).
>
> Basic. The Basic Level reflects marginal academic performance. Basic work indicates a partial

understanding and limited display of the skills included in the Pennsylvania Academic Content Standards. This work is approaching satisfactory performance, but has not been reached. There is a need for additional instructional opportunities and/or increased student academic commitment to achieve the Proficient Level (DRC, 2011, p. 218).

Below Basic. The Below Basic Level reflects inadequate academic performance. Below Basic work indicates little understanding and minimal display of the skills included in the Pennsylvania Academic Content Standards. There is a major need for additional instructional opportunities and/or increased student academic commitment to achieve the Proficient Level (DRC, 2011, p. 218).

The reliability of overall test scores in reading was strong with alpha coefficients at .91 for grade 3, and .89 for grades 4 and 5 (DRC, 2011). Decision consistency using the performance level indicators was evaluated using the Hanson and Brennan (1990) and Livingston and Lewis (1995) methods. According to Won-Chan Lee (2008), estimating decision consistency involves estimating both the true score and the observed score distributions. Both the Hanson and Brennan and Livingston and Lewis methods "employ a family of beta distributions for estimating the true score distribution" (p.

1). The results of these procedures yielded decision consistency coefficients of .76 - .79 in grades 3 - 5. Inter-rater agreement for open-ended responses in reading ranged from 73 – 85%.

Coefficients were not reported by the DRC regarding any type of validity; however, the DRC suggested that a strong link is shown between each item on the PSSA reading test and the assessment anchor it is intended to measure (DRC, 2011). They further state that the PSSA was carefully aligned to the PSSA assessment anchors followed by several content and bias reviews and field tests. Selected items had to pass rigorous reviews by the PDE for psychometric properties and content. Also, inter-item correlations show that reading items correlate more highly with themselves that than they do with other subjects such as math. This helps to show highlight evidence regarding the relationships, both internal and external, between the components of the assessments.

National and Locally-Generated Benchmark Expectations

Given the importance of the decisions made with the DIBELS benchmarks, more information is needed to understand if the nationally-derived benchmarks created by the DIBELS system are providing the most accurate criterion for reading proficiency, particularly when applied at the state or school district level. The DIBELS benchmarks are calculated based on

performance with a nationally-normed standardized achievement test (Good et al. 2011b). Although this procedure may reflect reading proficiency at the national level, it may not accurately represent the standard of performance on a state assessment. This is because differences can exist between the level of proficiency from one state to the next (Kingsbury et al., 2003). Using a nationally normed assessment like the DIBELS benchmarks to monitor progress toward the end of the year proficiency assessment is complicated by these inter-state differences. The result may show a high number of false positives and false negatives when using the benchmark score to predict state test scores. Therefore, a different criterion that more accurately reflects local expectations may be needed.

This section discusses the advantages and rationale for the use of locally-generated benchmarks. The procedures utilized by the authors of DIBELS to generate their benchmark scores are presented. This presentation includes the design specifications for both the DIBELS benchmarks and the cut points for risk. The benchmark scores developed by the DIBELS system for students in grades 3 - 5 are also included. In addition, the calculation procedures for the locally-generated benchmarks are also discussed. A description of a logistic

regression procedure (Silberglitt & Hintze, 2005) used in this study is presented as well as a rationale for its use.

DIBELS Benchmark Goals and Cut Points for Risk

According to the authors, the DIBELS benchmark goals are empirically derived, criterion-referenced target scores that represent adequate reading progress (Good et al., 2011b). "A benchmark goal indicates a level of skill where the student is likely to achieve the next DIBELS benchmark goal or reading outcome" (Good et al., 2011b, p. 46). In other words, students who are able to reach the benchmark expectation are more likely to achieve later reading goals when effective, evidence-supported teaching methods are employed.

Use of the benchmark goals correspond with universal screening procedures assessed with DIBELS. Students are assessed in the beginning, middle and end of each school year (Good, Kaminski, Dewey et al., 2011). The ORF scores earned by students on the DIBELS are then compared to benchmarks expectations, or cut scores, which indicate the likelihood that a student will or will not reach subsequent reading goals. In addition to the benchmarks, DIBELS also provides *Cut Points for Risk.* These cut points indicate the likelihood that an at-risk student will reach subsequent reading goals without the aid of supplemental instruction.

Students who score at or above the benchmark are identified as *Likely to Need Core Support* or *Low-Risk* (Good, Kaminski, Dewey et al., 2011). According to the DIBELS system, the odds that a student who achieves a score in this area will achieve early literacy goals are 80 - 90% (Good et al., 2011b). Students who fall below the benchmark goal fall in one of two categories. Students that fall below the benchmark expectation but above the cut point for risk are identified as *Some-Risk* or *Likely to Need Strategic Support*. The odds that a student who falls in this category will reach early reading goals are between 40 - 60%. Students who fall below the cut point for risk are titled *At-Risk* or *Likely to Need Intensive Support*. The odds that a student identified as *At-Risk* will reach early literacy goals are between 10 - 20%. The Benchmark Goals and cut points for risk for students in grades 3 - 5 in ORF are displayed in Table 1.

Table 1

DIBELS Benchmark Goals and Cut Points for Risk

Grade	Beginning	Middle	End
Grade 3			
Benchmark Goal	70	86	100
Cut Point for Risk	55	68	80
Grade 4			
Benchmark Goal	90	103	115
Cut Point for Risk	70	79	95
Grade 5			
Benchmark Goal	111	120	130
Cut Point for Risk	96	101	105

Calculation procedures. In the development of their benchmark goals, Good et al. (2011b) use a step-by-step process which will enable the user to monitor student progress toward successful reading outcomes. Students who achieve the benchmark goal in the beginning of the school year will likely meet the benchmark goal in the middle of the year. Students who are then able to reach the goal in the middle of the year will have excellent odds in achieving the target at the end of the year. Finally, students who have reached the end of the year benchmark goal will likely show adequate performance on any number of external measures of reading skills. According to the authors of DIBELS:

> Our fundamental logic for developing the benchmark goals and cut points for risk was to begin with the external outcome goal and work backward in that step-by-step system. We first obtained an external criterion measure (the GRADE Total Test Raw Score) at the end of the year with a level of performance that would represent adequate reading skills. Next we specified the benchmark goal and cut point for risk on the end of the year DIELS composite score with respect to the end-of-year external criterion. Then, using the DIBELS composite score as the internal criterion, we established the benchmark goals and cut

> points for risk on the middle-of-year DIBELS composite score. Finally, we established the benchmark goals and cut points for risk on the beginning-of-year DIBELS composite score using the middle-of-year DIBELS composite score as an internal criterion (p. 48).

For the purposes of their benchmarking procedures, the authors used a total test raw score on the GRADE as their external criterion for success in their benchmark development (Good et al., 2011b). Based on expectations for basic and proficient performance denoted by the National Assessment of Educational Progress (NAEP), the authors determined that a total test raw scores at or above the 40th percentile is a valid estimate of successful reading skills. Similarly, a total test raw score on the GRADE that was equal to or below the 20th percentile was utilized as the external criterion for the cut point for risk. According to the authors, this criterion should reflect Below Basic performance as denoted by the NAEP.

Primary design features. The primary design feature for the benchmark goal was to identify a skill level where students who score above the benchmark will reach successful reading goals at a minimum of 80% - 90% of the time (Good, Kaminski, Dewey, et al., 2011). Their primary design specification for the cut points for risk was a level of skill

where students who score below the cut point will likely reach appropriate reading goals only 10% - 20% of the time (Good et al., 2011b). Students who score between the benchmark goal and the cut point for risk have approximately even odds (40% - 60%) in achieving appropriate reading goals.

Secondary design features. In addition to the primary design features, the authors also considered secondary design features (Good et al., 2011b). They considered an equi-percentile method where they examined the marginal percents of the students who fell in each score level. They conducted this examination so that a consistent percentage of students fell in both the predictor criterion and the predicted criterion. For example, their findings show that 73% of third-grade students assessed fell above the 40th percentile on the GRADE. To remain consistent, they set the end of the year benchmark so that 73% of students fell above the benchmark score. By following this design specification, they were able to maintain a consistent level of performance from one indicator to the next. In addition, the authors used logistic regression to predict the odds of earning a score on the criterion measure that was equal to or above the benchmark expectation based on the predictor score (Good et al., 2011b).

Other considerations. After consideration of the primary and secondary design specifications, four other factors were

given consideration in the development of the benchmark expectations (Good et al., 2011b). First, they conducted a visual inspection of student performance on scatter plots. This method was used to create consistent goals which would highlight that students who score at or above the benchmark on the predictor variable will also be at or above the benchmark on the predicted variable. Similarly, students who fell below the cut point for risk on the predictor variable would generally show the same performance on the outcome measure. Second, receiver operator characteristics curve (ROC) analysis was also analyzed for Area Under Curve (AUC). Large AUC is preferable to ensure a balance between sensitivity and specificity. By balancing sensitivity and specificity, the authors sought to maximize accurate predictions with the benchmark goals and to minimize false positives and false negatives. Through their analysis, they found AUC range of .88 - .96 for the grade 3 - 5 benchmark expectations. Other metrics were analyzed for sensitivity (.71 - .84), specificity (.88 - .91), NPP (.87 - .92), PPP (.71 - .83), percent accurate classification (.83 - .89), and kappa (.55 - .75). Third, the knowledge gained from previous incarnations of the DIBELS as well as previously developed benchmark goals was also considered when developing the current benchmark goals and cut points for risk. Finally, the authors used their

overall professional judgment when incorporating the information from the primary, secondary, and other design specifications. Thus the benchmark goals are the result of the authors best compromises for all of the methods discussed.

Rationale for Developing Locally-Generated Benchmarks

The step-by-step process employed in the development of DIBELS is designed to carefully balance the number of false positives and false negatives (Good et al., 2011b). Once the process is completed the resultant benchmarks reflect adequate general reading outcomes across the country. Unfortunately, as a tool which can be utilized on a national scale, the developers of DIBELS may have sacrificed diagnostic accuracy at the individual state, school district, or elementary school building level (Silberglitt, 2008). Because they validate their indicators with performance on a nationally-normed standardized achievement test, the result of the validation may not necessarily reflect a state assessment level of proficiency. This is especially relevant given that research has shown discrepancies between the level of difficulty from one state proficiency test to the next (Kingsbury et al., 2004). Kingsbury et al. (2004) found large discrepancies between the percentile scores needed on a nationally-normed test of achievement that can accurately predict performance on a state assessment. These inter-state differences occur

because each state addresses the requirements under NCLB differently. They develop their own standards of performance and create their own assessments to meet those standards. This individual approach to education has resulted in state assessments of varying levels of difficulty and standards of performance which are disparate.

Differences between state proficiency levels and tests can create difficulties when using a nationally-normed assessment like the DIBELS benchmarks to monitor student progress toward the end of the year proficiency assessment. Using a DIBELS benchmark that has not been validated with the state test may result in a high number of false positives and false negatives. Consequently, a student who earns an ORF score above DIBELS benchmark score may not meet proficiency in a state with a higher standard. Similarly, a student who earns a below benchmark ORF score may still meet proficiency in a state where the standard for proficiency is much lower. Because proficiency is relative from one state to another, a one-size-fits-all benchmark score may not be a sufficient predictor of state standards and expectations.

A particular difficulty arises when applying the DIBELS benchmarks for special education decision making within an RtI model. The authors developed the benchmarks to be used as an effective universal screening measure. Their purpose was to

create a level of performance that would ensure that students in need of additional support would be identified. To accomplish this goal, the authors maximized sensitivity and specificity in a way which raises the benchmark score to guarantee that more students would be identified deficient readers (Good et al., 2011b). This process ensures that fewer students would miss out on much needed interventions. Unfortunately, it also increases the number false positives or students identified as deficient who earn a proficient score on the state test (Howell, Hosp, & Kurns, 2008). Although appropriate for screening decisions, the use of inflated DIBIELS benchmarks as a determinate of an insufficient level of performance in a dual-discrepancy approach within an RtI model may both unfair and inappropriate. This is because it uses a standard of performance that that does not accurately represent the level needed to perform successfully on the state assessment.

These concerns do not suggest that the DIBELS benchmarks are without value. In the absence of a national curriculum or national assessment, measures like DIBELS may highlight what proficient reading looks like across the country. Because of the discrepancies from one state to the next, school districts should avoid simply employing a nationally-normed set of benchmarks that may not be connected to their local

expectations (Silberglitt, 2009). Using a locally-generated benchmark score linked to a state assessment provides an internally consistent set of criteria which can be used to judge student progress (Silberglitt & Hintze, 2005). These benchmarks offer an educational standard that is linked to a familiar comparison group of students who have similar educational experiences and demographic backgrounds (Stewart & Kaminksi, 2002). The use of a local comparisons group reduces the chance of bias, as students would not be unfairly compared to norm groups that do not reflect their cultural, ethnic, linguistic, or economic background. Depending on the procedures utilized, locally-generated may be a more accurate standard which can be used for high-stakes decision making (Hintze & Silberglitt, 2005).

Locally-generated benchmarks appear to have a substantial potential for use in an RtI model. The connection between a locally-generated benchmark and state test performance creates a standard that is likely to be more reflective of local expectations than nationally-generated benchmarks. Once developed, these locally-generated benchmarks may be employed in universal screening procedures within Tier 1 or as a standard to which student growth can be compared at Tiers 2 and 3. An additional use of locally-generated benchmarks may be as a criterion of performance in high states special

educational determination decisions. In the following section, the calculation procedures for the development of locally-generated benchmarks are discussed. A logistic regression procedure recommended by Silberglitt and Hintze (2005) is presented along with a rational for its use.

Calculation Procedures for Locally-Generated Benchmarks

Silberglitt and Hintze (2005) evaluated different statistical procedures for generating local target scores. The purpose was to determine which procedure was most accurate and useful for developing and applying CBM benchmark scores. The sample for the study included 2,191 students from the St. Croix River Education District (SCRED), an education consortium comprised of five member school districts that reside in rural and suburban settings. The ethnic background of the sample was predominately White (95.3%) with 2.1% Native American, 1.4% Asian, 0.6% Hispanic, and .06% Black not of Hispanic Origin. Fifty-three percent of all participants were male and 26% received free or reduced lunches.

The participants completed the Minnesota Comprehensive Assessment (MCS) in the spring of grade 3 and at least one ORF benchmark assessment in grades 1 – 3 (Silberglitt & Hintze, 2005). During the ORF benchmarks assessments, students were administered standard reading assessment passages as developed by Howe and Shinn (as cited in Silberglitt & Hintze, 2005).

Three ORF probes were given at each grade level in the fall, winter, and spring, with the median score selected to represent the overall level of ORF at each time period.

The correlations between the ORF and MCA were moderate to strong ranging from .47 in the winter of grade 1 to .71 in the spring of grade 3. These findings are consistent with previous research that found coefficients near .70 between ORF and state assessment scores (Atkins & Cummings, 2011; Buck & Torgesen, 2003; Crawford, Tindal, & Steiber, 2001; Good, Simmons, & Kame'enui, 2001; Hintze & Silberglitt, 2005; McGlinchey & Hixon, 2004; Merino & Beckman, 2010; Schilling et al., 2005; Shaw & Shaw, 2002; Silberglitt et al., 2006); Stage & Jacobson, 2001). The correlations consistently became progressively stronger the closer in time ORF and the MCA were assessed.

Four different statistical procedures were analyzed including discriminant analysis, an equi-percentile method, logistic regression, and ROC curve analysis (Silberglitt & Hintze, 2005). Discriminant analysis was defined by the authors as a procedure that would be used to predict the likelihood of membership in a group by evaluating a collection of variables in a population. The equi-percentile method compared the percentage of students who meet proficiency on an outcome measure to the percentage of students on the predictor

measure. The goal is to equalize the percentages of students on the predictor variable to the outcome measure. The benchmark score selected at the point where the percentages of students on both assessments are equal. Logistic regression is a statistical procedure used for categorical dependent variables. It allows the user to employ one or more quantitative or categorical variables to predict a categorical outcome (Silberglitt, 2008). It is best used with CBM to predict a dichotomized outcome such as proficient or below proficient. Finally, ROC curve analysis method plots the sensitivity and specificity of a predictor for all values of the target score. Target scores are subsequently selected by using research supported judgments to balance desirable levels of sensitivity and specificity.

With each procedure the authors analyzed the number of students who did or did not reach proficiency on the MCA using ORF scores as the predictor variable (Silberglitt & Hintze, 2005). For each cut score calculation method, diagnostic accuracy statistics were calculated including PPP, NPP, sensitivity, and specificity. In addition Silberglitt and Hintze (2005) calculated kappa, the standard error of kappa, and a phi Coefficient. Table 2 illustrates the diagnostic accuracy statistics for each cut score method.

Table 2

A Comparison of Cut Scores and Diagnostic Statistics for Predicting Success on the MCAS Using CBM-R Scores at Spring of Grades 1 through 3

Grade	Method	Cut Score	PPP	NPP	Sen	Spec	Kappa	Standard Error of Kappa	Phi
Grade 3									
	DA	106	.691	.832	.801	.733	.525	.022	.529
	EQUI	100	.730	.807	.736	.802	.537	.022	.538
	LR	98	.748	.803	.722	.823	.547	.022	.548
	ROC	107	.685	.835	.801	.733	.522	.022	.527
Grade 2									
	DA	88	.660	.823	.778	.720	.485	.023	.490
	EQUI	82	.709	.811	.737	.788	.523	.023	.523
	LR	79	.729	.794	.696	.819	.519	.023	.519
	ROC	90	.654	.832	.797	.706	.486	.023	.494
Grade 1									
	DA	54	.837	.837	.808	.646	.428	.025	.445
	EQUI	45	.797	.797	.702	.768	.465	.025	.466
	LR	43	.786	.786	.671	.795	.467	.025	.467
	ROC	49	.812	.812	.751	.706	.442	.025	.447

Note. DA = Discriminant Analysis; EQUI = Equipercentile Method; LR = Logistic Regression; ROC = Receiver Operating Characteristic Curve Analysis; Sen = sensitivity; Spec = specificity. Reprinted from "Formative Assessment Using CBM-Cut Scores to Track Progress Toward Success on State-Mandated Achievement Tests: A Comparison of Methods," by B. Silberglitt and J. Hintze, 2005, *Journal of Psychoeducational Assessment, 23*, p. 318. Copyright 2005 by Sage Publications. Reprinted with permission.

Based on their comparison, Silberglitt and Hintze (2005) recommended logistic regression, ROC curve analysis, or a combination of both methods for establishing standards and creating benchmark scores. Logistic regression produced the lowest cut scores but also yielded the highest percentage of overall accuracy of prediction as well as the highest values for PPP. In turn, logistic regression will create less false positives but will identify fewer at-risk students. ROC curve analysis yielded the highest cut scores which consequently produced the highest values for NPP. This is important to

note because higher cut scores will reduce false negatives and will help to ensure that more students will be identified as at-risk.

Hintze and Silberglitt (2005) suggested that the selection of a method for calculating locally-generated benchmark scores depends on the purpose for which the cut score is to be used. Given the flexibility using ROC curve analysis, cut scores can be manipulated to maximize their sensitivity and specificity. The user can then adjust the cut score to make appropriate screening decisions which would ensure more students are identified as needing supports. For classification purposes such as special education eligibility determination, however, logistic regression may be a more appropriate method. This is because logistic regression produces cut scores which are efficient, accurate, and consistent. When compared to other methods of calculation, logistic regression consistently shows the highest percentage of overall accuracy of prediction (Hintze & Silberglitt, 2005; Silberglitt & Hintze, 2005). Because it maximizes the true positives and does not produce a benchmark score which is inflated to ensure more students receive early intervening services (Hintze & Silberglitt, 2005).

For these reasons, logistic regression was chosen to develop the benchmarks in this study. This was to ensure that

a benchmark score would be developed which would accurately predict performance on the PSSA. The benchmarks would be developed using a procedure similar to the reverse validated step-by-step process utilized by the DIBELS system (Good et al., 2011b). Using logistic regression, the spring ORF score were linked to performance on the PSSA to create the spring benchmark score. This benchmark became the criterion used to create the winter benchmark score. Fall ORF performance were connected to the generated winter benchmark score. Using this method ensured that a consistent set of procedures were used for both the DIBELS benchmarks and the locally-generated benchmarks.

Summary

The literature relevant to the components of a response to intervention model was discussed in this chapter with a focus on the assessment procedures and decision making practices utilized in the RtI process. This discussion included procedures for universal screening with CBM with ORF. The research relevant to CBM demonstrated a moderate to strong correlation between ORF and general reading achievement. Similar, findings between ORF and performance on state assessments, including the PSSA, were identified. A detailed discussion of benchmark expectations for identifying students in need of additional reading support was presented. In

particular, the benchmark expectations provided by the DIBELS System were discussed along with a rationale for utilizing locally-generated benchmarks. Methods for calculating both the DIBELS benchmarks and local ORF benchmarks were included in this chapter.

CHAPTER III

METHOD AND PROCEDURES

Introduction

The validity of locally-generated benchmark expectations to predict performance on the Pennsylvania System of School Assessment (PSSA) is examined in this study. The generated benchmarks, once compared to the benchmarks created by the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) system, will help to resolve whether the DIBELS benchmarks are sufficient for predicting performance on the PSSA or whether locally-generated benchmarks should be considered.

The methods and procedures used to answer the research questions included in this study are discussed in this chapter. A description of the study sites that participated was provided including location, demographics, and instructional characteristics. In addition, descriptions of the instruments used in this study are presented. The sample, derived from the two study sites, was divided into two separate groups using a random selection procedure. This procedure was conducted to create a comparison group that can be used to cross-validate the benchmark scores. A logistic regression procedure was utilized to develop the locally-generated benchmark scores and the statistical analyses,

including descriptive statistics and diagnostic accuracy statistics were presented.

Design

This study explored whether the DIBELS benchmark goals or locally-derived benchmark goals have a stronger relationship and predictive validity with performance on the PSSA. DIBELS Oral Reading Fluency (ORF) data and state proficiency assessment data were collected from two school districts in the Commonwealth of Pennsylvania. The data were analyzed to uncover any statistically significant mean differences on DIBELS benchmarks between the two participating schools. This procedure was conducted to determine whether the data from the participating study sites could be combined to develop one set of benchmark scores or whether sets of benchmark scores would be generated for each study site. The resultant benchmarks were compared to the benchmarks created by the DIBELS system. This would help determine whether the DIBELS benchmarks are sufficient for predicting performance on the PSSA or whether locally-generated benchmarks should be considered.

Student data were divided into two groups based upon the study site where they were enrolled. In both study site data sets, student scores were randomly divided into two separate groups allowing for an independent data set to be used for cross-validation purposes. Cross-validation was conducted to

ensure that the generated benchmark scores would generalize to subsequent cohorts of students (Jenkins, Hudson, & Johnson, 2007). In the absence of cross-validation, benchmark scores applied to subsequent cohorts of students may not achieve the same level of predictive accuracy as what was initially obtained.

Population

Archival and anonymous data from the 2010 - 2011 school year were examined in this study. The data were collected from two school districts located in central Pennsylvania. The names of the participant school districts are withheld to ensure anonymity.

Study Site 1

The data were collected from a rural school district in central Pennsylvania. In the 2010 - 2011 school year, approximately 2,500 students were served by this school district (National Center for Education Statistics, 2012). Less than 1% of students were identified as English Language Learners (ELL). Fourteen percent of students in this district were serviced with an Individualized Education Program (IEP). This percentage special education enrollment is consistent with the Pennsylvania state average of 14.87% (Pennsylvania Department of Education, 2011). This school district contains four neighborhood elementary school buildings with similar

size and demographics. There was a 14.5 to 1 average student-to-teacher ratio at the elementary school buildings. Approximately 18% of students received free and reduced lunches (NCES, 2012). The median family income within this school district is approximately 60,000 dollars per year (Proximity, 2013).

In the 2010 - 2011 school year, 538 students were enrolled in grades 3 - 5. White/Non-Hispanic students comprised the majority of the student population at 91.17% (NCES, 2012). Approximately, 1% of the student population was made up of Black/Non-Hispanic students with 5.65% of the student population reporting to be of Hispanic descent. Less than 1% identified as an Asian/Pacific Islander.

Study Site 2

Data were collected from a suburban school district located in central Pennsylvania. In the 2010 - 2011 school year, approximately 3,500 students were enrolled in this school district (NCES, 2012). Similar with Study Site 1, less than 1% of students were identified as ELL. Twelve percent of students received services through an IEP, which is lower than the state average of 14.87% (NCES, 2012; Pennsylvania Department of Education, 2011). The elementary school system is divided into three distinctive levels including Early Childhood (Kindergarten and Grade 1), Primary (grades 2 and 3)

and Intermediate (grades 4 and 5). Each of these educational levels is located within the same school building. There was a 14.5:1 student-to-teacher ratio at this study site (NCES, 2012). Approximately 15% of students received free and reduced lunches. The median family income within this school district is approximately 84,000 dollars per year (Proximity, 2013).

During the 2010 - 2011 school year, 793 students were enrolled in grades 3 - 5. Approximately 79% of these students identified as White/Non-Hispanic with 10.6% who identified as Asian/Pacific Islander (NCES, 2012). Hispanic students comprised 4.6% of the student population with 5.5% Black/Non-Hispanic students.

Sample

Inclusion Criteria

All archival and anonymous records for students enrolled in grades 3 - 5 at both participating school districts during the 2010 - 2011 school year were examined. All students who completed each of the DIBELS benchmark assessments and the PSSA were selected.

Exclusion Criteria

Exclusion criteria were solely based upon the availability of the data to be analyzed. Students whose records were incomplete were excluded from this study.

Incomplete records may include: missing one or more of the DIBELS benchmark scores, missing PSSA scores, or incomplete demographic information including sex, race, and socio-economic status as defined by free-and-reduced-lunch status.

Assignment

This study represents a sample of convenience as only archival data were analyzed. Student data were divided into two groups based upon the study site where they were enrolled. In both study sites, student scores were randomly divided into two separate groups. This allowed for an independent data set that could be used to cross-validate both the DIBELS and locally-generated benchmark scores.

Measurement

Dependent Variable

The dependent variable in this study was student performance on the PSSA. The PSSA is a criterion-referenced test used to assess students' progress toward the acquisition of skills related to the Pennsylvania State academic standards (Pennsylvania Department of Education, 2010). It is also used as a measure of a school district's success developing programs which enable students to reach the standards. Students in Pennsylvania are assessed in reading in grades 3 - 8 and grade 11. Reading scores in grades 3 - 5 were analyzed in this study. Scores on the PSSA are reported as scaled

scores. The proficiency categories Below Basic, Basic, Proficient, and Advanced were utilized according to the cut scores developed by the Pennsylvania Department of Education.

The internal consistency of the overall test scores in reading is strong with Cronbach's Alpha coefficients above .9 (Data Recognition Corporation, 2011). The reliability or consistency of decisions across the four performance level descriptors was also evaluated. Across all subject area, the consistency of the descriptors ranged from .76 - .79 in grades 3 - 5. Inter-rater reliability for the open ended responses ranged from .73 - .85 (DRC, 2011).

Validity coefficients were not reported by the Data Recognition Corporation (DRC; Shapiro, 2006) however, the DRC (2011) suggested that evidence of the validity of the PSSA is demonstrated by the strong link between each item on the reading test and the assessment anchor it is intended to measure. The DRC also stated that the PSSA was carefully aligned to the PSSA assessment anchors followed by several content and bias reviews and field tests. Selected items had to pass rigorous reviews by the Pennsylvania Department of Education for psychometric properties and content. The DRC also reported that inter-correlations between reading items are stronger than correlations with reading items and items within other subject areas such as math. This helped to

highlight evidence regarding the relationships, both intern and external, between the components of the assessment.

Independent Variable

The independent variable in this study is ORF benchmark scores on the DIBELS. DIBELS ORF is a curriculum-based measure (CBM) where students are asked to read aloud from a passage for 1 minute (Good et al., 2011a). During the fall, winter, and spring assessment periods 3 1–minute passages are administered to each student. The median number of words the student read correctly in 1 minute is used as the primary metric to determine student performance (Shinn, 2008). This procedure is followed per the standardized administration procedures of the DIBELS system.

The test retest reliability of ORF in grades 3 - 5 ranges from .93 - .97 (Powell-Smith et al., 2011). Test-retest reliability was not assessed for ORF. Inter-rater reliability ORF was high at .99 for all three grade levels.

Criterion-related validly measuring the extent to which the student performance on the Group Reading Assessment and Diagnostic Evaluation was predicted based on their DIBELS performance (Powell-Smith et al., 2011). Predictive validly coefficients ranged from ORF from .66 - .77 in grades 3 - 5 (Powell et al., 2011). Concurrent validity coefficients for ORF ranged from .65 - .74.

Procedure

Existing archival data were examined in this study. DIBELS data, PSSA scores, and demographic information including sex, race, and free-and-reduced-lunch status were gathered by the school psychologist at Study Site 1 and by the director of psychological services at Study Site 2. Each student record received an alphanumerical code (Participant 1, Participant 2, etc.) to ensure that these data were provided to the primary researcher with all identifying information removed. At no time was the primary researcher given access to personally-identifiable information.

In each grade level, student data were divided into two groups based upon the study site where they were enrolled. Additionally, both data sets were then separated into two groups, the Benchmark group and Comparison group. This division provided an independent data set that was used to cross-validate both the DIBELS and locally-generated benchmark scores. To accomplish the separation, students were separated into four sub-groups as determined by their performance level on the PSSA (Below Basic, Basic, Proficient, and Advanced). Random case selection was conducted using SPSS in each of the four sub-groups. Approximately fifty percent of the participants in each sub-group were selected to create the Comparison group. In the event of an odd number of students

in a subgroup, the Comparison group received the higher number of participants. The division of the participants in the manner was done to ensure that all ranges of student reading skills were represented.

Locally-generated benchmarks were developed using a logistic regression procedure (Hintze & Silberglitt, 2005; Silberglitt, 2008; Silberglitt & Hintze, 2005). Logistic regression allows quantitative or categorical data to be analyzed in a way which will predict a categorical outcome (Neter, Kutner, Nachtsheim, & Wasserman, 1996). The assumptions for logistic regression include a dichotomous dependent variable, large sample size, non-multicollinearity, and interdependence of errors (Hosmer & Lemeshow, 2000; Neter et al., 1996.

Following the logistic regression, the ORF score which produced the highest overall percentage of correct PSSA predictions was selected as the benchmark score. This score was selected as it maximizes the number of true positives (Hintze & Silberglitt, 2005). The benchmark scores at each time period (beginning, middle, and end) within each grade level were selected in this manner. Similar to the step-by-step reverse validated procedure utilized by the DIBELS system, the spring ORF score was linked to performance on the PSSA to create the spring benchmark score. This benchmark

then became the criterion used to create the winter benchmark score. Fall ORF was then connected to the generated winter benchmark score. Using this method ensured that a consistent set of procedures was used for both the DIBELS benchmarks and the locally-generated benchmarks. Furthermore, Hintze and Silberglitt (2005) suggested that this method produces reliable and valid benchmark expectations, particularly when logistic regression is used.

Data Analyses

The following statistical analyses were used to examine each research question in this study. The Statistical Package for the Social Sciences (SPSS), Version 20 was used to analyze the data. Each of the research questions along with the hypotheses, statistical methods, and assumptions necessary to answer the questions are illustrated in Table 2.

Research Question 1

Are there statistically significant mean differences on DIBELS benchmarks between the two participating schools? It is hypothesized that differences will not exist between the benchmark scores generated for the participating school districts. This hypothesis was made because the study sites share relatively similar racial, sex, and socio-economic demographic characteristics. In addition, analogous

percentages of special education students and English Language Learners are present in both study sites.

Research Question 2

What are the correlations between the fall, winter, and spring DIBELS ORF scores and performance on the PSSA in grades 3 - 5? Separate correlations will be calculated for each grade and each assessment period (fall, winter, and spring).

The analysis of this research question is dependent upon the results of Research Question 1. Benchmark scores from the schools will be generated and correlated with PSSA in grades 3 – 5. If the two participating schools' benchmark scores are significant different, as determined from the analyses associated with Research Question 1, then separate benchmarks will be generated for both study sites. If significant differences are not identified, the data for both study sites will be combined to generate one set of scores.

Consistent with previous research (Baker et al., 2008; Deno, Mirkin, & Chiang, 1982; Hintz & Silberglitt, 2005; Hosp & Fuchs, 2005; Jenkins & Jewell, 1993; Shinn et al., 1992), it is hypothesized that ORF at each grade level will demonstrate moderate to strong correlations with the results of the PSSA.

Research Question 3

What are the locally-generated benchmark scores in the fall, winter, and spring in grades 3 - 5? Logistic regression

was used to calculate the locally-generated benchmarks. This procedure was chosen because the cut score in a range of student ORF scores that produces the highest percentage of correct predictions of the PSSA is selected and utilized as part of the analysis with logistic regression.

If significant differences are identified between the benchmark scores of participating sites, then separate benchmarks will be generated for both study sites. If significant differences are not identified, the data for both study sites will be combined to generate one set of benchmark scores.

It is hypothesized that the locally-generated benchmarks will be lower than those created by the DIBELS system. This is because the DIBELS benchmarks are designed to carefully balance the number of false positives and false negatives in an attempt to ensure that more students are identified as in need of additional supports (Good et al., 2011b). This inflated cut score, useful for screening purposes, sacrifices the accuracy of the prediction of PSSA proficiency. In addition, the logistic regression procedure used in this study maximizes the percentage of true positives only and produces a benchmark score that is not artificially inflated but maximizes the prediction accuracy on the PSSA. According to Hintze and Silberglitt (2005), logistic regression typically

produces cut scores that are lower than other methods of calculation.

Research Question 4

Are the locally-generated benchmarks able to predict PSSA proficiency with significantly greater accuracy than the DIBELS benchmarks? Additionally, are measures of diagnostic accuracy (sensitivity, specificity, PPP, and NPP) significantly different based on the derivation of the benchmarks?

The analysis of this research question is dependent upon the results of Research Question 1. If significant differences are identified between the benchmark scores, then benchmarks scores generated for both study sites will be compared separately with the DIBELS benchmark scores. If significant differences are not identified, then the benchmark scores for both study sites will be combined to generate one set of benchmark scores that will be compared to the DIBELS benchmarks.

Descriptive statistics and diagnostic accuracy statistics including specificity, sensitivity, PPP, and negative predicative power were calculated to show the percentage of students who passed and failed the PSSA based on their scores in ORF. Total accuracy percentages, kappa and phi of both benchmarks were calculated to determine which more accurately

predicts performance on the PSSA. Significant differences between the total accuracy percentages and values for kappa were measured through the use of z-score tests.

It was hypothesized that significant differences would be identified between the locally-developed benchmarks and the DIBELS-generated benchmarks in their ability to reliably predict PSSA performance. It is further hypothesized that the locally-generated benchmarks would more accurately predict PSSA performance. In addition, significant differences will be present between the diagnostic accuracy statistics for both sets of benchmarks. These hypotheses were suggested for two reasons. First, the DIBELS benchmarks are designed to carefully balance the number of false positives and false negatives in order to create an inflated cut score that ensures more students are identified as in need of additional supports (Good et al., 2011b). This inflated score, however, will likely produce a less accurate prediction of PSSA proficiency. Second, locally-generated benchmarks developed by Ferchalk et al. (2010) more accurately predicted proficiency on the PSSA than DIBELS-generated benchmarks. Similar findings were predicted for this study.

Summary

The methods and procedures used to answer four research questions evaluating the differences between locally-generated

benchmarks and the benchmarks provided by the DIBELS system were discussed in this chapter. An explanation of the purpose and design of the study was provided along with a description of the population, sample, and method of assignment. The two instruments used in the study were discussed and the as well as the reliability and validity for both measures. The procedures used for calculating the locally-generated benchmarks were detailed along with the statistical analyses used to answer the research questions.

Table 3

Research Questions, Hypotheses, Variables, Statistical Analyses, and Statistical Assumptions

Research Questions	Hypotheses	Variables	Statistical Analyses	Statistical Assumptions
1. Are there statistically significant mean differences on the DIBELS benchmarks between the two participating schools?	Differences will not exist because the study sites share similar characteristics and percentages of special education and ELL.	Fall, Winter, and Spring ORF scores and PSSA scaled scores.	t-tests	1. Interval or ratio data 2. Normal distributions 3. Sample Size 4. Equal variances
2. What are the correlations between the fall, winter, and spring DIBELS ORF scores and performance on the PSSA in grades 3-5?	ORF at each grade level will demonstrate moderate to strong correlation with the results of the PSSA.	Fall, Winter, and Spring ORF scores and PSSA scaled scores.	Pearson Correlation	1. Interval or ratio data 2. Normal distributions 3. Minimal outliers 4. Equal variances
3. What are the locally-generated benchmark scores in the fall, winter, and spring in grades 3 – 5?	The score which maximizes the overall predictive accuracy will be selected as the benchmark score.	Fall, Winter, and Spring ORF scores and PSSA scaled scores.	Logistic Regression	1. Dichotomous Dependent variable 2. Large sample size 3. Non-multicollinearity 4. Interdependence of errors
4. Are the locally-generated benchmarks able to predict PSSA proficiency with significantly greater accuracy than the DIBELS benchmarks? Additionally, are measures of diagnostic accuracy (sensitivity, specificity, PPP, and NPP) significantly different based on the derivation of the benchmarks?	The locally-generated benchmark will more accurately predict PSSA performance than those generated by the DIBELS system.	Local ORF benchmark scores and the DIBELS ORF Benchmark scores	z-score Tests Descriptive Statistics	1. Interval data 2. Normality 3. Equal variances

CHAPTER IV

DATA AND ANALYSIS

The use and validity of locally-generated Oral Reading Fluency (ORF) benchmark expectations to predict performance on the Pennsylvania System of School Assessment (PSSA) was examined in this study. The generated benchmarks, once compared to the benchmarks created by the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) system, will help to resolve whether the DIBELS benchmarks are sufficient for predicting performance on the PSSA or whether locally-generated benchmarks should be considered.

Four research questions that investigative the relationship between ORF and the PSSA are further examined in this chapter. The results for each of the four research questions are presented. The first research question was posed to determine if the data collected from both study sties could be aggregated into one overall data set. The results of this analysis would determine how subsequent research questions would be analyzed.

Complications

In both study sites, only students who completed the PSSA reading test were included in this study. This excluded students who may have completed the Pennsylvania Alternate System of Assessment or the PSSA – Modified version. Their

exclusion from this study may have skewed the results as it eliminated some students with disabilities from the analysis. In addition, several students were removed from both data sets due to missing information. In Study Site 1, six students were removed from the data sets at each grade level as they were missing either ORF or PSSA data. The missing data in Study Site 1 accounted for approximately 4% of the total sample. In Study Site 2, 20 students were removed from the grade 3 data set, 15 were removed from the grade 4 data set and 15 were removed from the grade 5 data set. The missing data in Study Site 2 accounted for approximately 7% of the total sample. The small percentage of missing data in both study sites is not believed to have greatly affected the results of this study.

The missing information in this study was most likely the result of student transience. Students who moved in to the district in the middle of the year would have missed one or more of the DIBELS assessments. Similarly, students who were enrolled in the beginning of the year but have moved to another district before the end of the year would have missed one or more of the DIBELS assessments as well as the PSSA. Additional explanations include English Language Learners who may not have participated in the PSSA if this is their first

year of education in the United States and students who were absent during the PSSA administration period.

Research Question 1

Are there statistically significant mean differences on DIBELS benchmarks between the two participating schools? It was hypothesized that differences would not exist between the scores for the participating school districts. This hypothesis was made because the study sites share relatively similar racial, sex, and socio-economic demographic characteristics. In addition, analogous percentages of special education students and English Language Learners are present in both study sites.

Grade 3 descriptive statistics are displayed in Table 4. In Study Site 1, mean scores of 77.61, 100.02, and 109.95 were obtained in the fall, winter, and spring, respectively. Study Site 2 yielded mean scores of 94.00, 111.41, and 126.72 in the fall winter and spring. Standard deviations in both study sites ranged from 36.26 – 37.74.

The normality of the data was first assessed using visual inspection of the frequency distributions. Each of the frequency distributions of the ORF scores examined approximated a normal bell-shaped curve. Skewness and kurtosis statistics were also analyzed to determine normality. A range of -1 to 1 was chosen to represent an acceptable level

of skewness (Breakwell, 2006; Greer, Dunlap, Hunter, & Berman, 2006). In both study sites, values for skewness ranged from .10 - .45. These scores fall within the acceptable range indicating that the ORF distributions were symmetrical. Kurtosis statistics were also analyzed in both study sites. A range of -3 to 3 to was pre-selected to represent an acceptable level of kurtosis (Anastasi, 1982; Gaur & Gaur, 2006). In both study sites, values for kurtosis ranged from -.42 – .15 and fell within the acceptable range. This indicates that the ORF data were normally distributed.

Table 4

Descriptive Statistics for Grade 3 ORF and PSSA Scores

	Range	*X*	*SD*	Skewness	Kurtosis
Study Site 1[a]					
Fall	8 – 185	77.61	37.01	.45	-.21
Winter	21 – 208	100.02	36.26	.41	-.01
Spring	22 – 235	109.95	37.74	.30	-.14
PSSA	1000 - 1681	1318.70	140.27	.00	-.18
Study Site 2[b]					
Fall	10 – 210	94.00	37.62	.29	-.42
Winter	23 – 212	111.41	36.35	.10	-.26
Spring	36 – 256	126.72	36.35	.13	.15
PSSA	1035 - 1942	1376.86	149.76	.27	1.23

[a]*n* = 171. [b]*n* = 260.

The mean PSSA score of Study Site 1 was of 1318.70 with a standard deviation of 140.27. A mean score of 1376.86 and standard deviation of 149.76 was found in Study Site 2. Visual inspection of the frequency distribution approximately normally distributed PSSA data. Additionally, values for

skewness and kurtosis indicated that the PSSA score distributions approximate normality.

Descriptive statistics for Grades 4 are reported in Table 5. In Grade 4, mean ORF scores of 95.97, 115.27, and 127.26 were obtained in the fall, winter, and spring, respectively, at Study Site 1. In Study Site 2, mean ORF scores of 117.11 (fall), 133.38 (winter), and 144.74 (spring) were obtained. Standard deviations were similar between both study sites ranging from 32.30 – 37.41. Values for skewness ranged from –.01 - .38. Values for Kurtosis ranged from –.51 - .06. Skewness and kurtosis values indicated that the frequency distributions approximated a normal curve. This was confirmed though visual examination of the frequency distributions.

Table 5

Descriptive Statistics for Grade 4 ORF and PSSA Scores

	Range	*X*	*SD*	Skewness	Kurtosis
Study Site 1[a]					
Fall	24-188	95.97	38.90	.38	-.51
Winter	40-198	115.27	37.42	.22	-.57
Spring	52-233	127.26	37.04	.37	-.44
PSSA	830-1801	1306.86	162.69	.11	.41
Study Site 2[b]					
Fall	18-235	117.11	37.41	.27	-.15
Winter	30-236	133.38	32.30	-.01	.33
Spring	36-247	144.74	33.06	.02	.06
PSSA	910-1891	1426.71	176.41	.02	.30

[a]*n* = 141. [b]*n* = 246.

A mean PSSA score of 1306.86 with a standard deviation of 162.69 were produced in Study Site 1. In Study Site 2, a mean score of 1426.71 and standard deviation of 176.41 were

identified in Study Site 2. Visual inspection of the frequency distributions and analysis of the skewness and kurtosis statistics indicated that the PSSA score distributions approximate normality.

Similar findings were identified in Grade 5 as shown in Table 6. Mean fall, winter, and spring ORF scores of 122.21, 139.09, and 143.23, respectively, were obtained for Study Site 1. In Study Site 2, mean scores of 128.18 (fall), 144.16 (winter), and 153.56 (spring) were identified. Standard deviations were again similar across both study sites ranging from 31.80 – 33.28. Skewness values ranged from –.07 - .08 with Kurtosis ranging from –.51 - .02. Both skewness and kurtosis statistics indicated that the frequency distributions approximated a normal curve. This was confirmed though visual examination of the frequency distributions.

Table 6

Descriptive Statistics for Grade 5 ORF and PSSA Scores

	Range	*X*	*SD*	Skewness	Kurtosis
Study Site 1[a]					
Fall	57-200	122.21	32.51	.07	-.51
Winter	70-228	139.09	32.50	.08	-.46
Spring	68-231	143.23	32.01	-.04	-.24
PSSA	739-1701	1307.79	174.27	-.04	.15
Study Site 2[b]					
Fall	20-227	128.18	33.28	-.07	-.15
Winter	34-234	144.16	32.67	.06	-.11
Spring	54-244	153.56	31.80	-.02	.02
PSSA	850-1933	1398.82	187.46	-.08	.14

[a]*n* = 180. [b]*n* = 256.

A mean PSSA score of 1307.79 with a standard deviation of 174.27 were produced in Study Site 1. A mean score of 1398.82 and standard deviation of 187.46 was found in Study Site 2. Visual inspection of the frequency distributions and analysis of the skewness and kurtosis statistics indicated that the PSSA score distributions approximated normality.

To determine if significant differences were present between the study sites, a series of independent samples t-test were conducted. The t-tests were carried out between both study sites at each of the three benchmark assessment assessments and at each grade level. The assumptions for t-tests, including the use of interval or ratio data, normal distributions in each data set, and appropriate sample sizes were met. Equal variances were assured by an examination of the descriptive statistics. In addition, the results of Levene's Test for equality of variances indicated that the distributions had approximately equivalent amounts of variability between scores.

The results of the *t*-tests are reported in Tables 7 - 9. Significant differences were identified between nearly all of the benchmark scores except for the fall, $t(434) = -1.86$, $p = .063$, and winter, $t(434) = -1.60$, $p = .111$, of fifth grade. During these two assessment periods, mean differences of only 5 words correct per minute (wcpm) were identified.

Table 7

Grade 3 Independent Samples t-tests for Fall, Winter, and Spring DIBELS ORF

	Levene's Test		*t*-test for Equality of Means			
	F	*P*	*t*	*df*	Mean Difference	Std. Error Difference
Fall	.26	.61	-4.45*	429	-16.39	3.68
Winter	.33	.56	-3.19*	429	-11.39	3.58
Spring	.43	.52	-4.61*	429	-16.76	3.63

* $p < .01$, two tailed.

Table 8

Grade 4 Independent Samples t-tests for Fall, Winter, and Spring DIBELS ORF

	Levene's Test		*t*-test for Equality of Means			
	F	*P*	*t*	*df*	Mean Difference	Std. Error Difference
Fall	.55	.46	-5.60*	391	-22.14	3.96
Winter	6.03	.02	-5.07*	391	-18.11	3.58
Spring	2.85	.09	-4.85*	391	-17.48	3.61

* $p < .01$, two tailed.

Table 9

Grade 5 Independent Samples t-tests for Fall, Winter, and Spring DIBELS ORF

	Levene's Test		*t*-test for Equality of Means			
	F	*P*	*t*	*df*	Mean Difference	Std. Error Difference
Fall	.14	.71	-1.86	434	-5.97	3.21
Winter	.29	.59	-1.60	434	-5.07	3.17
Spring	.09	.76	-3.33*	434	-10.31	3.10

* $p < .01$, two tailed.

Between the remaining benchmark scores, mean differences ranged from 10 to 22 wcpm. In grade 3, significant differences were identified between each of the fall, winter, and spring benchmark scores the benchmark scores, $t(429) = -3.19 - (-4.61)$, $p < .01$. In grade 4, significant differences were also found between benchmark scores in each of the three

time periods assessed, $t(391) = -4.85 - (-5.60)$, $p < .01$. Benchmark scores in the spring of grade 5 were also significant $t(434) = -3.33$, $p < .01$. It is noted that study Site 2 consistently earned higher scores than Study Site 1.

The differences between both study sites were also reflected in PSSA proficiency rates as shown in Table 10. In this table, the percentages of students in each PSSA category in both study sites are reported. In addition, student performance levels were dichotomized into either a pass or fail group. Students who passed the PSSA scored Proficient or Advanced on the PSSA. Students who scored in the Basic or Below Basic range were considered to have failed the PSSA. In grade 3, Study Site 2 11% more students passed the PSSA than Study Site 1. The difference was 23% higher in grade 4 and 16% higher in grade 5.

Table 10

Percentage of Students in Each PSSA Category

Site	*N*	Bel	Bas	Pro	Adv	Fail (Bel/Bas)	Pass (Pro/Adv)
Grade 3							
Site 1	171	14%	12%	56%	19%	26%	74%
Site 2	260	9%	6%	52%	34%	15%	85%
Grade 4							
Site 1	147	12%	26%	50%	12%	38%	62%
Site 2	246	6%	9%	48%	38%	15%	85%
Grade 5							
Site 1	180	13%	30%	43%	14%	43%	57%
Site 2	256	7%	20%	43%	30%	27%	73%

Note. Bel = Below Basic, Bas = Basic, Pro = Proficient, Adv = Advanced.

Independent samples *t*-tests were conducted to determine if the differences between PSSA scores earned by the two study were significant. This set of *t*-tests was carried out between both study sites at each grade level. The assumptions for t-tests, including the use of interval or ratio data, normal distributions in each data set, and appropriate sample sizes were met. Equal variances were assured by an examination of the descriptive statistics. In addition, the results of Levene's Test for equality of variances indicated that the distributions had approximately equal amounts of variability.

The results of the t-tests are reported in Table 11. In grade 3, the mean differences between PSSA scores in both study sites was significant, $t(429) = -4.04$, $p < .01$. Significant differences were also found in grade 4, $t(391) = -6.71$, $p < .01$, and grade 5, $t(434) = -5.13$, $p < .01$. It is noted that study Site 2 consistently earned higher scaled scores on the PSSA than Study Site 1.

Table 11

Independent Samples t-tests for Grades 3 - 5 PSSA Scaled Scores

	Levene's Test		*t*-test for Equality of Means			
	F	*p*	*t*	*df*	Mean Difference	Std. Error Difference
Grade 3	.04	.84	-4.04*	429	-58.16	14.38
Grade 4	.40	.53	-6.71*	391	-119.85	17.87
Grade 5	1.72	.19	-5.13*	434	-91.03	17.49

* $p < .01$, two tailed.

Given these results, the null hypothesis is rejected. Significant differences did exist between the majority of the ORF scores for both study sites as indicated by the t-test results. The scores earned on the PSSA in both study sites add additional evidence to support these findings. As result, the data collected from both study sites will be analyzed separately for each of the three remaining research questions.

Research Question 2

What is the correlation between DIBELS ORF and performance on the PSSA in grades 3 - 5? Consistent with previous research (Baker et al., 2008; Deno et al., 1982; Hintz & Silberglitt, 2005; Hosp & Fuchs, 2005; Jenkins & Jewell, 1993; Shinn et al., 1992), it is hypothesized that ORF at each grade level will demonstrate moderate to strong correlations with the results of the PSSA.

Given the results of the previous research question, Pearson correlations were calculated between ORF at each assessment period (fall, winter, and spring) and the PSSA at both study sites separately. The study sites were not separated into the Benchmark and Comparison groups for this research question as in the subsequent two research questions. The data collected were instead analyzed as two complete study sites. The assumptions for Pearson correlation, including the use of interval or ratio data, normal distributions in each

data set, appropriate sample sizes, equal variances and minimal outliers were met as determined by an examination of the descriptive statistics.

The strength of the correlations was evaluated using ranges recommended by Evans (1996). Correlations ranging from .40 to .59 were defined as moderate while correlations ranging from .60 - .79 were determined to be strong. Coefficients above .70 were defined as very strong.

The correlations are reported in Tables 12 – 14. Correlations between the fall, winter, and spring ORF measures were very strong, ranging from .90 - .95 across all three grade levels. Each correlation was statistically significant at the .01 level. Correlations between ORF and the PSSA were strongest in grade 3 in Study Site 1 (.69 - .70). In Study Site 2, correlations in grade 3 were also strong ranging from .64 - .66. In both study sites, the strength of the correlations decreased slightly but remained in the moderate to strong ranges in subsequent years. In Grade 4, correlations in Study Site 1 ranged from .61 - .64 and from .54 - .57 in Study Site 2. Grade 5 correlations in Study Site 1 ranged from .57 - .60 with correlations of .61 - .62 in Study Site 2. Each correlation was statistically significant at the .01 level.

Table 12

Correlation Matrix for Grade 3 ORF and PSSA

Measure	1	2	3	4
Study Site 1				
1. DIBELS Fall	-			
2. DIBELS Winter	.93*	-		
3. DIBELS Spring	.90*	.93*	-	
4. PSSA	.69*	.70*	.69*	-
Study Site 2				
1. DIBELS Fall	-			
2. DIBELS Winter	.91*	-		
3. DIBELS Spring	.90*	.93*	-	
4. PSSA	.64*	.66*	.66*	-

*$p < .01$, two-tailed.

Table 13

Correlation Matrix for Grade 4 ORF and PSSA

Measure	1	2	3	4
Study Site 1				
1. DIBELS Fall	-			
2. DIBELS Winter	.95*	-		
3. DIBELS Spring	.92*	.94*	-	
4. PSSA	.64*	.63*	.61*	-
Study Site 2				
1. DIBELS Fall	-			
2. DIBELS Winter	.93*	-		
3. DIBELS Spring	.90*	.92*	-	
4. PSSA	.56*	.57*	.54*	-

*$p < .01$, two-tailed.

Table 14

Correlation Matrix for Grade 5 ORF and PSSA

Measure	1	2	3	4
Study Site 1				
1. DIBELS Fall	-			
2. DIBELS Winter	.92*	-		
3. DIBELS Spring	.90*	.92*	-	
4. PSSA	.60*	.57*	.60*	-
Study Site 2				
1. DIBELS Fall	-			
2. DIBELS Winter	.94*	-		
3. DIBELS Spring	.91*	.94*	-	
4. PSSA	.60*	.62*	.61*	-

*$p < .01$, two-tailed.

It was hypothesized that the correlations found between ORF and the PSSA analyzed in this study would be consistent with those found in previous research. This hypothesis is supported by the correlations found in this study. Similar correlations were obtained in previous research studies between measures of reading fluency and performances on both general reading assessments and state assessments (Buck & Torgesen, 2003; Schilling, et at., 2005; Crawford, Tindal, & Steiber, 2001; Deno, Mirkin, & Chiang, 1982; Ferchalk, Cogan-Ferchalk, & Richardson, 2012; Good, Simmons, & Kame'enui, 2001; Hintze & Silbergiltt, 2005; Hosp & Fuchs, 2005; Jenkins & Jewell, 1993; Keller-Margulis, Shapiro, & Hintze; 2008; Shaw & Shaw, 2002; McGlinchey & Hixon, 2004 ; Shapiro, et al. 2006; Shapiro, Solari, & Petscher, 2008; Silberglitt, et al. 2006).

Fisher's z transformations were used to analyze the differences between the correlation coefficients obtained in both study sites. An average of the correlations produced in the fall, winter, and spring were analyzed in each grade level. In addition, Fisher transformations were utilized to determine if the average correlations obtained in the current study are similar to those found by Silberglitt et al. (2006). The authors of that study analyzed the relationship between measures of ORF and performances on the Minnesota

Comprehensive Assessment in Reading (MCA). The MCA was not assessed in grade 4. Consequently, no correlations were calculated by Silberglitt et al. at this grade level. The results of the Fisher's z transformations are reported in Table 15.

Table 15

Fisher's z Transformations Comparing Coefficients Between Study Site 1, Study Site 2 and Coefficients Found in Silberglitt et al. (2006)

			Fisher's z		
	n	*r*	1	2	3
Grade 3					
1 Study Site 1	171	.69	-	0.73	0.24
2 Study Site 2	260	.65	-0.73	-	-0.83
3 Silberglitt et al. (2006)	3165	.68	-0.24	0.83	-
Grade 4					
1 Study Site 1	141	.63	-	0.88	-
2 Study Site 2	246	.56	-0.88	-	-
3 Silberglitt et al. (2006)	-	-	-	-	-
Grade 5					
1 Study Site 1	180	.59	-	-0.32	-1.27
2 Study Site 2	256	.61	0.32	-	-1.02
3 Silberglitt et al. (2006)	3283	.65	1.27	1.02	-

Note. Adapted from "Relationship of Reading Fluency Assessment Data with State Accountability Test Scores: A Longitudinal Comparison of Grade Levels," by B. Silberglitt, M. K. Burns, N. H. Madyun, and K. E. Lail, 2006, *Psychology in the Schools*, *43*, p. 531. Copyright 2006 by Wiley Periodicals, Inc.
*$p < .05$, two-tailed.

An alpha level of .05 was selected to determine if the differences between the coefficients were significant. The results of the Fisher transformations show that the average correlations obtained for Study Site 1 and Study Site 2 did not show significant differences at any grade level assessed. Similarly, the average correlation coefficients in grades 3 and 5 obtained in this study were not significantly different

from those found by Silberglitt et al. (2006). Consistent with the hypothesis, the lack of difference between the coefficients show that the correlations obtained in the present study are consistent with those found in previous research.

Research Question 3

What are the locally-generated benchmark scores in the fall, winter, and spring in grades 3 - 5? Logistic regression was used to calculate the locally-generated benchmarks. This procedure was chosen because the cut score in a range of student ORF scores that produces the highest percentage of correct predictions of the PSSA is selected and utilized as part of the analysis with logistic regression.

It was hypothesized that that the locally-generated benchmarks would be lower than those created by the DIBELS system. The developer of DIBELS designed their benchmarks so that a higher number of students would be identified as at-risk (Good et al., 2011b). This inflated cut score, useful for screening purposes, sacrifices the accuracy of the prediction of PSSA proficiency (Hintze & Silberglitt, 2005). In addition, the logistic regression used in this study maximized the percentage of true positives only and produced a benchmark score that was not artificially inflated but maximized the prediction accuracy on the PSSA. According to

Hintze and Silberglitt (2005), logistic regression produces cut scores that are lower than other statistical methods of calculation.

The results of Research Question 1 indicated significant differences between the mean ORF scores in both study sites. Consequently, separate sets of benchmarks were generated for both study sites. In each grade level, within both study sites, data were separated into two groups: the Benchmark group used to generate the benchmark scores and the Comparison group to which the scores would be applied. To accomplish the separation, students were separated into four sub-groups as determined by their performance level on the PSSA (Below Basic, Basic, Proficient, and Advanced). Approximately, 50% of the participants in each sub-group were selected to create the Benchmark group and the remaining 50% comprised the Comparison group.

Tables 16 – 21 show the descriptive statistics for both the benchmark and Comparison group at each grade level and study site. An examination of the descriptive statistics in both study sites revealed that the data for the Benchmark and Comparison groups were evenly distributed. At each grade level, similar ranges and comparable mean scores and standard deviations were found in both groups. Similarly, the PSSA scaled scores within both the Benchmark and Comparison groups

yielded analogous score ranges and comparable means and standard deviations.

A series of independent samples t-tests were conducted between the Benchmark and Comparison groups at each time period assessed and at each of the three grade levels. No significant differences were identified between the ORF scores and the PSSA scores for the two groups. These findings suggest that students in both the Benchmark and Comparison groups earned relatively equivalent DIBELS ORF scores and PSSA scores across all assessment periods and across each grade level assessed. The similarity between the Benchmark and Comparison groups helps to confirm the appropriateness of conducting the cross-validation study of the locally-generated ORF cut scores.

Table 16

Grade 3 Study Site 1 Descriptive Statistics

	Range	*X*	*SD*	Skewness	Kurtosis
Benchmark[a]					
Fall	14-165	77.60	36.78	.52	-.23
Winter	24-192	98.78	34.37	.33	-.09
Spring	40-194	109.26	36.76	.43	-.48
PSSA	1017-1618	1323.94	134.75	.02	-.49
Comparison[b]					
Fall	8-185	77.62	37.45	.39	-.14
Winter	21-208	101.27	38.25	.45	.01
Spring	22-235	110.66	38.92	.20	.20
PSSA	1000-1681	1313.38	146.26	.00	.06

Note. Fall, Winter, and Spring data refer to ORF wcpm; PSSA data represent a scaled score.
[a]n = 86. [b]n = 85.

Table 17

Grade 3 Study Site 2 Descriptive Statistics

	Range	*X*	*SD*	Skewness	Kurtosis
Benchmark[a]					
Fall	10-210	94.00	39.66	.27	-.49
Winter	23-212	112.44	38.56	.11	-.35
Spring	36-256	126.76	38.64	.25	.54
PSSA	1035-1942	1384.85	164.88	.52	1.28
Comparison[b]					
Fall	16-181	93.79	35.91	.32	-.36
Winter	31-199	110.16	34.51	.08	-.24
Spring	49-205	126.65	34.52	-.04	-.58
PSSA	1035-1782	1368.72	134.92	-.26	.38

Note. Fall, Winter, and Spring data refer to ORF wcpm; PSSA data represent a scaled score.
[a]*n* = 128. [b]*n* = 129.

Table 18

Grade 4 Study Site 1 Descriptive Statistics

	Range	*X*	*SD*	Skewness	Kurtosis
Benchmark[a]					
Fall	24-182	95.09	38.17	.42	-.58
Winter	40-198	117.07	37.31	.29	-.65
Spring	52-205	127.39	36.35	.28	-.70
PSSA	1018-1801	1316.23	163.64	.56	.50
Comparison[b]					
Fall	29-188	94.84	39.90	.35	-.40
Winter	44-197	113.44	37.70	.17	-.48
Spring	62-233	127.12	37.98	.47	-.17
PSSA	830-1637	1297.36	162.30	-.36	.20

Note. Fall, Winter, and Spring data refer to ORF wcpm; PSSA data represent a scaled score.
[a]*n* = 74. [b]*n* = 73.

Table 19

Grade 4 Study Site 2 Descriptive Statistics

	Range	*X*	*SD*	Skewness	Kurtosis
Benchmark[a]					
Fall	47-235	117.38	39.47	.49	-.10
Winter	59-236	132.91	33.91	.22	.22
Spring	75-247	144.11	34.05	.27	.11
PSSA	977-1891	1421.18	176.35	.02	-.00
Comparison[b]					
Fall	18-198	116.84	35.36	-.05	-.28
Winter	30-209	133.85	30.71	-.31	.53
Spring	36-220	145.37	32.14	-.27	.08
PSSA	910-1891	1432.33	177.01	.03	.68

Note. Fall, Winter, and Spring data refer to ORF wcpm; PSSA data represent a scaled score.
[a]*n* = 124. [b]*n* = 122.

Table 20

Grade 5 Study Site 1 Descriptive Statistics

	Range	*X*	*SD*	Skewness	Kurtosis
Benchmark[a]					
Fall	59-199	122.64	32.82	.06	-.57
Winter	74-205	138.27	32.48	-.01	-.94
Spring	68-220	142.56	31.33	-.14	-.26
PSSA	739-1701	1304.18	185.48	-.18	.34
Comparison[b]					
Fall	57-200	121.78	32.36	.09	-.40
Winter	70-228	139.93	32.67	.17	-.33
Spring	73-231	143.92	32.87	.04	-.20
PSSA	963-1701	1311.48	162.99	.06	-.26

Note. Fall, Winter, and Spring data refer to ORF wcpm; PSSA data represent a scaled score.
[a]*n* = 91. [b]*n* = 89.

Table 21

Grade 5 Study Site 2 Descriptive Statistics

	Range	*X*	*SD*	Skewness	Kurtosis
Benchmark[a]					
Fall	66-199	128.75	31.73	.11	-.97
Winter	84-220	144.60	32.23	.38	-.79
Spring	87-225	153.98	30.47	.19	-.71
PSSA	942-1835	1403.11	178.14	-.03	-.43
Comparison[b]					
Fall	20-227	127.60	34.90	-.20	.38
Winter	34-234	143.72	33.24	-.23	.50
Spring	54-244	153.13	33.24	-.17	.53
PSSA	850-1933	1394.46	197.10	-.10	.46

Note. Fall, Winter, and Spring data refer to ORF wcpm; PSSA data represent a scaled score.
[a]*n* = 129. [b]*n* = 127.

A procedure using logistic regression (Silberglitt, 2008) was utilized to determine the benchmark score. Upon completion of logistic regression, the ORF score that maximized the overall accuracy of the prediction of PSSA proficiency was selected as the benchmark score at each time period (beginning, middle, and end) at each grade level. The assumptions for calculating logistic regression including dichotomous dependent variable, large sample size, non-

multicollinearity, and interdependence of errors were met. The dependent variable, PSSA proficiency, was dichotomized as pass/fail for this study. Interdependence of errors was assured as each set of ORF data collected were not from a dependent samples design. The independent variables did not show multicollinearity as only one independent variable was included for each logistic regression procedure calculated.

A step-by-step reverse validated procedure was utilized with logistic regression to calculate the benchmark scores. Logistic regression was first conducted in the spring of the school year using ORF as the continuous independent variable with pass/fail performance on the PSSA as a dichotomous dependent variable. After analyzing the results, the cut score produced by the logistic regression was chosen as the locally-generated benchmark for the spring ORF data. This locally-generated benchmark score was then applied to the spring ORF data set and used to dichotomize the student scores as pass/fail. The winter ORF scores were then used as the continuous independent variable with pass/fail performance on the spring ORF measure as the dichotomous dependent variable. The cut-score produced by logistic regression during this analysis was then applied to the winter ORF data set to dichotomize the data. This process was then repeated with the fall ORF scores serving as the continuous independent variable

with winter pass/fail ORF performance as a dichotomous dependent variable.

Table 22 depicts the logistic regression analyses for both Study Site 1 and Study Site 2. The significance of the regression coefficients was evaluated using the Wald chi square statistic. The relationship between the continuous independent variables and the dichotomous dependent variables were significant on each time period calculated in Study Site 1. This suggests that ORF performance in each time period significantly predicted pass/fail performance on the PSSA.

Table 22

Logistic Regression Analysis for Study Site 1

	B	*SE*	Wald's X^2	e^{β}
Grade 3[a]				
Fall to Winter	.21	.07	10.11*	1.24
Winter to Spring	.18	.05	10.73*	1.20
Spring to PSSA	.06	.02	16.03*	1.06
Grade 4[b]				
Fall to Winter	.18	.05	10.50*	1.19
Winter to Spring	.13	.04	13.34*	1.14
Spring to PSSA	.04	.01	14.12*	1.04
Grade 5[c]				
Fall to Winter	.13	.03	17.69*	1.14
Winter to Spring	.13	.03	18.03*	1.14
Spring to PSSA	.03	.01	12.52*	1.03

Note. $df = 1$.
[a]$n = 86$. [b]$n = 74$. [c]$n = 91$.
*$p < .05$.

Hosmer – Lemesmeshow goodness of fit tests for Study Site 1 are depicted in Table 23. The results yielded X^2 that were not significant ($p > .05$) in all of the data periods assessed. This indicates that the model was a good fit to the data.

Table 23

Study Site 1 Hosmer – Lemeshow Tests for Goodness of Fit

	X^2	*df*	*p*
Grade 3[a]			
Fall to Winter	1.02	7	.99
Winter to Spring	1.42	7	.99
Spring to PSSA	8.01	7	.33
Grade 4[b]			
Fall to Winter	2.69	8	.95
Winter to Spring	1.12	8	1.00
Spring to PSSA	4.33	8	.83
Grade 5[c]			
Fall to Winter	3.40	8	.91
Winter to Spring	2.05	8	.98
Spring to PSSA	11.13	8	.20

[a]*n* = 86. [b]*n* = 74. [c]*n* = 91.
*$p < .05$

Logistic Regression analysis for Study Site 2 is illustrated in Table 24. In Study Site 2, the relationships between the ORF scores and the predicted dichotomous outcome in grade 3, the spring of grade 4, and all of grade 5 were significant. This pattern was not continued in the fall and winter of 4th grade.

Table 24

Logistic Regression Analysis for Study Site 2

	B	*SE*	Wald's X^2	e^{β}
Grade 3[a]				
Fall to Winter	.15	.04	12.07*	1.16
Winter to Spring	.17	.05	11.88*	1.18
Spring to PSSA	.08	.02	21.24*	1.08
Grade 4[b]				
Fall to Winter	.37	.22	2.82	1.45
Winter to Spring	.17	.07	5.44*	1.18
Spring to PSSA	.38	.01	13.34*	1.04
Grade 5[c]				
Fall to Winter	.18	.05	14.34*	1.19
Winter to Spring	.18	.05	16.52*	1.20
Spring to PSSA	.05	.01	22.54*	1.05

Note. *df* = 1.
[a]*n* = 128. [b]*n* = 124. [c]*n* = 129.
*$p < .05$.

Table 25 illustrates the Hosmer – Lemesmeshow tests for Study Site 2. The results of these tests yielded X^2 that were not significant (p >.05) in all of the data time periods assessed excluding the fall to winter of grade 5, $X^2(7) = 32.11$, $p < .05$. This indicates that the model was a good fit to the data in this study site.

Table 25

Study Site 2 Hosmer – Lemeshow Tests for Goodness of Fit.

	X^2	*df*	*p*
Grade 3[a]			
Fall to Winter	.71	8	1.00
Winter to Spring	.16	9	1.00
Spring to PSSA	4.09	8	.85
Grade 4[b]			
Fall to Winter	.00	8	1.00
Winter to Spring	.11	8	1.00
Spring to PSSA	4.77	8	.78
Grade 5[c]			
Fall to Winter	32.11	7	.00*
Winter to Spring	2.67	8	.95
Spring to PSSA	13.91	8	.08

[a]*n* = 128. [b]*n* = 124. [c]*n* = 129.
*P < .05

The locally-generated benchmarks developed using logistic regression as well as the DIBELS-generated benchmarks are illustrated in Table 26. In grade 3 at Study Site 1, the locally-generated benchmarks were 19 – 26 wcpm lower than the DIBELS benchmarks. The locally-generated benchmarks in Study Site 2 were 19 – 29 wcpm lower than the benchmarks derived from the DIBELS system. In grade 4 at Study Site 1, the locally-generated benchmarks were 9 – 21 wcpm lower than the DIBELS benchmarks. In Study Site 2, the differences between

the locally-generated benchmarks and the DIBELS benchmarks ranged from 29 – 36 wcpm.

Table 26

DIBELS Benchmark Goals and Locally-Generated Benchmarks

	Fall	Winter	Spring
Grade 3			
DIBELS Benchmark	70	86	100
Study Site 1	44	67	79
Study Site 2	41	60	81
Grade 4			
DIBELS Benchmark	90	103	115
Study Site 1	69	91	106
Study Site 2	54	68	86
Grade 5			
DIBELS Benchmark	111	120	130
Study Site 1	111	124	133
Study Site 2	92	112	127

Note. Data indicate ORF wcpm.

The differences between locally-generated and DIBELS-generated benchmarks in Grade 5 were slight. In Study Site 1, the locally-generated benchmarks were equal to or higher than those of the DIBELS System. The fall locally-generated benchmark score in Study Site 1 was equal to the DIBELS benchmark. In the winter, Study Site 1 produced a score that was 4 points higher than the DIBELS benchmark. In the spring, the Study Site 1 locally-generated benchmark was again higher (3 wcpm) than the DIBELS benchmark.

The Differences between the locally-generated benchmarks and the DIBELS benchmarks in Study Site 2 diminished in grade 5. In the fall, the difference between the two benchmarks was 19 wcpm. During the winter assessment, the difference between the two benchmark cutscores decreased to only 8 wcpm. The

Study Site 2 spring locally-generated benchmark was only 3 points lower than the benchmark produced by the DIBELS system.

The strength of the differences between the locally-generated benchmarks and DIBELS benchmarks was measured through a comparison of the scores in relation to the standard deviations of the data set from which they were derived. It was proposed that if the differences between the benchmarks scores exceeded the standard deviation for the data set then the scores would be considered appreciatively different. Differences that did to exceed the standard deviation would help to suggest that the benchmark scores were relatively similar. Tables 27 and 28 include the benchmark scores, differences and standard deviations in both study sites across all three grade levels.

Table 27

Differences Between Local and DIBELS Benchmark Goals in Study Site 1

	Local	DIBELS	Difference	*SD*
Grade 3[a]				
Fall	44	70	26	36.78
Winter	67	86	19	34.37
Spring	79	100	21	36.76
Grade 4[b]				
Fall	69	90	21	38.17
Winter	91	103	12	37.31
Spring	106	115	9	36.35
Grade 5[c]				
Fall	111	111	0	32.82
Winter	124	120	-4	32.48
Spring	133	130	-3	31.33

Note. Data indicate ORF wcpm.
[a]n = 86. [b]n = 74. [c]n = 91.

Table 28

Differences Between Local and DIBELS Benchmark Goals in Study Site 2

	Local	DIBELS	Difference	*SD*
Grade 3[a]				
Fall	41	70	29	39.66
Winter	60	86	26	38.56
Spring	81	100	19	38.64
Grade 4[b]				
Fall	54	90	36	39.47
Winter	68	103	35	33.91
Spring	86	115	29	34.05
Grade 5[c]				
Fall	92	111	19	31.73
Winter	112	120	8	32.23
Spring	127	130	3	30.47

Note. Data indicate ORF wcpm.
[a]n = 128. [b]n = 124. [c]n = 129.

The differences between the locally-generated benchmarks and the DIBELS benchmarks did not exceed the standard deviation during any time period or grade level in Study Site 1. Similarly, the differences between two sets of benchmarks did not exceed the standard deviation in Study Site 2 except in the winter of 4th grade. Therefore, the differences between the benchmarks are generally not considered to be substantial.

The purpose of this research question was to determine what are the locally-generated benchmarks using performance on the PSSA as the criterion for successful reading. This purpose was met using logistic regression in both study sites across all three grade levels. The hypothesis for this research question stated that the locally-generated benchmark would be lower than those generated by the DIBELS system. This hypothesis was not confirmed. Although differences were

present, the differences primarily did not exceed the standard deviation of the data sets from which the benchmarks were developed. Therefore, the null hypothesis is accepted.

Research Question 4

Are the locally-generated benchmarks able to predict PSSA proficiency with significantly greater accuracy than the DIBELS benchmarks? Additionally, are measures of diagnostic accuracy (sensitivity, specificity, PPP, and NPP) significantly different based on the derivation of the benchmarks?

It was hypothesized that significant differences will be identified between the locally-developed benchmarks and the DIBELS-generated benchmarks in their ability to predict PSSA performance. It is further hypothesized that the locally-generated benchmarks will more accurately predict PSSA performance. In addition, significant differences will be present between the diagnostic accuracy statistics for both sets of benchmarks. These hypotheses are suggested for two reasons. First, the creators of DIBELS went to great lengths to balance the percentages of false positives and false negatives when they developed their benchmark scores. The result of the procedures produced an elevated score that ensures more students are identified as at-risk and in need of supplementary reading interventions (Good et al., 2011b).

This inflated score, however, will likely produce a less accurate prediction of PSSA proficiency. Second, locally-generated benchmarks developed by Ferchalk et al. (2010) more accurately predicted proficiency on the PSSA than DIBELS-generated benchmarks. Similar findings are predicted for this study.

Diagnostic accuracy statistics including specificity, sensitivity, positive predictive power (PPP), and negative predictive power (NPP) were calculated to show the percentage of students who were accurately predicted to pass or fail the PSSA based on their scores in ORF. For the purpose of this analysis, specificity referred to the percentage of students who failed the PSSA and were accurately predicted to fail based on their ORF score. Sensitivity referred to the percentage of students who passed the PSSA and were accurately predicted to pass. PPP describes the percentage of students predicted to fail based on their ORF score who actually failed the PSSA. NPP portrays the percentage of students predicted to pass the PSSA based on their ORF score who actually passed the PSSA. In addition, the total percentage of accurate predictions was calculated along with kappa, standard error of kappa and phi to determine if locally-generated benchmarks or DIBELS benchmarks more accurately predict performance on the PSSA. Each of these statistics was calculated for both the

DIBELS benchmarks and the locally-generated cut scores. The diagnostic accuracy statistics for both study sites are displayed in Tables 29 and 30.

Table 29

Diagnostic Accuracy Statistics for Study Site 1 Locally-Generated Benchmarks and DIBELS Benchmarks

	Cut Score	Sen	Spec	PPP	NPP	Total %	Kappa	Standard Error of Kappa	Phi
Grade 3[a]									
Site 1 - F	44	0.68	0.95	0.83	0.90	0.88	0.67	0.10	0.68
DIBELS – F	70	0.91	0.78	0.59	0.96	0.81	0.58	0.09	0.61
Site 1 - W	67	0.59	0.97	0.87	0.87	0.87	0.62	0.11	0.64
DIBELS – W	86	0.86	0.84	0.66	0.95	0.85	0.64	0.09	0.65
Site 1 - S	79	0.64	0.94	0.78	0.88	0.86	0.61	0.11	0.61
DIBELS – S	100	0.86	0.78	0.58	0.94	0.80	0.55	0.10	0.58
Grade 4[b]									
Site 1 - F	69	0.63	0.91	0.81	0.81	0.81	0.57	0.10	0.58
DIBELS – F	90	0.89	0.78	0.71	0.92	0.82	0.64	0.09	0.65
Site 1 - W	91	0.59	0.91	0.80	0.79	0.79	0.54	0.11	0.55
DIBELS – W	103	0.81	0.83	0.73	0.88	0.82	0.63	0.09	0.63
Site 1 - S	106	0.70	0.93	0.86	0.84	0.85	0.66	0.09	0.67
DIBELS – S	115	0.81	0.83	0.73	0.88	0.82	0.63	0.09	0.63
Grade 5[c]									
Site 1 - F	111	0.61	0.80	0.70	0.73	0.72	0.42	0.10	0.42
DIBELS – F	111	0.61	0.80	0.70	0.73	0.72	0.42	0.10	0.42
Site 1 - W	124	0.58	0.78	0.67	0.71	0.70	0.37	0.10	0.37
DIBELS – W	120	0.50	0.80	0.66	0.68	0.67	0.31	0.11	0.32
Site 1 - S	133	0.63	0.86	0.77	0.76	0.76	0.51	0.09	0.51
DIBELS – S	130	0.58	0.90	0.81	0.74	0.76	0.50	0.10	0.52

Note. F = fall; W = winter; S = spring; Sen = sensitivity; Spec = specificity; NPP = negative predictive power; PPP = positive predictive power.
[a]n = 85. [b]n = 73. [c]n = 89.

Grade 3

In grade 3 within both study sites, the locally generated benchmarks produced higher levels of specificity (.94 - .99) than did the DIBELS benchmarks (.78 - .88), suggesting a more accurate prediction of true positives. The locally-generated benchmarks also produced higher levels of PPP (.78 - .90) than

the DIBELS benchmarks (.41 -.66). This would suggest than students who were predicted to fail the PSSA based on their ORF score had a 78 – 90% chance of failing the PSSA when using the locally-generated benchmark. Analysis of sensitivity and NPP indicates that the DIBELS benchmarks (sensitivity = .71 - .91; NPP = .94 - .96) outperformed the locally-generated benchmarks (sensitivity = .47 - .68; NPP = .87 - .90).

Table 30

Diagnostic Accuracy Statistics for Study Site 2 Locally-Generated Benchmarks and DIBELS Benchmarks

	Cut Score	Sen	Spec	PPP	NPP	Total %	Kappa	Standard Error of Kappa	Phi
Grade 3[a]									
Site 2 - F	41	0.47	0.99	0.89	0.93	0.92	0.58	0.13	0.61
DIBELS – F	70	0.76	0.83	0.41	0.96	0.82	0.43	0.11	0.47
Site 2 - W	60	0.53	0.99	0.90	0.93	0.93	0.63	0.12	0.66
DIBELS – W	86	0.71	0.88	0.46	0.95	0.85	0.47	0.11	0.49
Site 2 - S	81	0.47	0.98	0.80	0.92	0.91	0.55	0.13	0.57
DIBELS – S	100	0.76	0.83	0.41	0.96	0.82	0.43	0.11	0.47
Grade 4[b]									
Site 2 - F	54	0.18	0.99	0.75	0.88	0.88	0.25	0.18	0.33
DIBELS – F	90	0.59	0.83	0.36	0.93	0.80	0.33	0.12	0.34
Site 2 - W	68	0.18	0.99	0.75	0.88	0.88	0.25	0.18	0.33
DIBELS – W	103	0.29	0.90	0.33	0.89	0.82	0.21	0.15	0.21
Site 2 - S	86	0.18	1.00	1.00	0.88	0.89	0.27	0.18	0.40
DIBELS – S	115	0.41	0.86	0.32	0.90	0.80	0.24	0.14	0.24
Grade 5[c]									
Site 2 - F	92	0.29	0.95	0.67	0.79	0.77	0.29	0.12	0.33
DIBELS – F	111	0.74	0.83	0.61	0.90	0.80	0.53	0.08	0.53
Site 2 - W	112	0.35	0.95	0.71	0.80	0.79	0.36	0.11	0.39
DIBELS – W	120	0.53	0.89	0.64	0.84	0.80	0.45	0.10	0.45
Site 2 - S	127	0.50	0.94	0.74	0.84	0.82	0.49	0.10	0.50
DIBELS – S	130	0.53	0.91	0.69	0.84	0.81	0.48	0.10	0.49

Note. F = fall; W = winter; S = spring; Sen = sensitivity; Spec = specificity; NPP = negative predictive power; PPP = positive predictive power.
[a]*n* = 129. [b]*n* = 122, [c]*n* = 127.

In both study sites, the locally-generated benchmarks showed the highest levels of overall prediction accuracy as

illustrated by the overall accuracy percentage, kappa, and phi. The total percentage of accurate predictions ranged from .86 - .93 for the locally-generated benchmarks from both study sites. The total percentage of accurate predictions for the DIBELS benchmarks at both study sites ranged from .82 - .85. Values for both kappa and phi for the locally-generated benchmarks ranged from .55 - .68 indicating substantial agreement between ORF and PSSA proficiency. The DIBELS generated benchmarks produced slightly weaker agreement between ORF and the PSSA as evidenced by the values for both kappa and phi (.43 - .65).

Grade 4

In grade 4 within both study sites, the locally generated benchmarks continued to produce higher levels of Specificity (.91 – 1.00) than did the DIBELS benchmarks (.78 - .90). Higher levels of PPP were also produced by the locally-generated benchmarks (.75 - 1.00) than the DIBELS benchmarks (.32 -.73). The DIBELS benchmarks produced higher levels of sensitivity than the locally-generated benchmarks in both study sites. In Study Site 1 the DIBELS benchmark yielded sensitivity ranging from .81 - .89 with the locally generated benchmark yielding scores ranging from .59 – 70. The DIBELS benchmarks also showed higher sensitivity than was found in Study Site 2; however, very low levels of sensitivity was

produced by both the locally-generated benchmarks (.18) and the DIBELS benchmarks (.29 - .59). NPP was also higher in the DIBELS benchmark (.88 - .93) than the locally-generated benchmarks (.79 - .88).

In Study Site 1, the DIBELS benchmarks showed slightly higher levels of overall prediction accuracy in the fall and winter than the locally-generated benchmark as illustrated by the overall accuracy percentage, kappa, and phi. During these two assessment periods the DIBELS benchmarks predicted PSSA proficiency accurately 82% of the time, compared to the 79 - 81% accuracy rate of the locally-generated benchmarks. In the spring, the locally-generated benchmark was 85% accurate when predicting PSSA proficiency than the 82% of the DIBELS benchmark. Values for kappa and phi for the locally-generated benchmarks ranged from .54 - .67 indicating substantial agreement between ORF and PSSA proficiency. The DIBELS benchmarks produced slightly stronger agreement between ORF and the PSSA as evidenced by kappa and phi (.63 - .65). In Study Site 2, the locally-generated benchmark yielded higher total accuracy percentages (.88 - .89), than the DIBELS benchmark (.80 - .82). Values of kappa and phi for the locally-generated benchmark ranged from .21 - .33. The values for kappa and phi for the DIBELS benchmarks ranged from .12 - .34.

Grade 5

In grade 5 within Study Site 1, the locally-generated benchmarks and the DIBELS benchmarks yielded nearly identical benchmarks. Consequently, very similar values for all of the diagnostic accuracy statistics were produced for both. Both the local and DIBELS benchmarks produced high levels of specificity (.78 - .90) with lower levels of sensitivity (.50 - .63). Both NPP and PPP were relatively equal with percentage ranges of .68 - .76 and .66 -.81 respectively. The accuracy prediction of both sets of benchmarks were also equivalent with total accuracy percentage ranging from .67 - .76, and kappa and phi ranging from .31 - .52.

In Study Site 2, the DIBELS benchmarks remained higher than the locally-generated benchmark thereby producing higher levels of sensitivity (.53 -.74) and NPP (.84 - .90) than locally-generated benchmark (sensitivity = .29 - .50; NPP = .79 - .84). The locally-generated benchmarks continued to show higher levels of specificity (.94 - .95) and PPP (.67 - .74) than the DIBELS benchmarks (specificity = .83 - .91, PPP = .61 - .69). The DIBELS benchmarks, however, generally produced higher values for total accuracy percentage (.80 - .81) and kappa and phi (.45 - .53) than the locally-generated

benchmarks (total accuracy percentage = .77 - .82; kappa and phi = .29 - .50).

Differences Between Benchmark Scores

To determine if significant differences existed between the benchmarks scores, z-score tests were calculated for both the total accuracy percentages and the values for kappa. These tests were analyzed between the diagnostic accuracy statistics between both study sites. The following 3 equations, suggested by Sheskin (2004), were used to calculate the differences between the z – scores in this study. Equation 1 illustrates the equation used to calculate the z-score test between the two accuracy percentages.

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (1)$$

Equation 2 shows the equation used in this study to calculate an approximation for the standard error of kappa.

$$\sigma_{K_p} = \frac{\sum E_{ij}}{\sqrt{n(n - \sum E_{ij})}} \quad (2)$$

Equation 3 depicts the algorithm, suggested by Sheskin, used to calculate significant differences between two values for kappa.

$$z = \frac{(K_1 - K_2)}{\sqrt{\sigma_{K_1}^2 + \sigma_{K_2}^2}} \quad (3)$$

Results of the z-score tests did not yield significant differences between the locally-generated benchmarks and the DIBELS benchmarks in any grade level within Study Site 1 as depicted in Table 31. The greatest differences were present between the locally-generated benchmarks and DIBELS benchmarks in grade 3. These differences, however, were not significant. The calculated z-scores in this grade level for both total accuracy percentage (z = .38 – 1.26, p = 1.96) and kappa (z = −0.11 - .67, p = 1.96) did not fall outside the selected critical value of 1.96 established a priori. Therefore, the null hypothesis was accepted.

Table 31

z-Score Tests for Differences Between DIBELS- and Locally-Generated Benchmark Diagnostic Accuracy Statistics in Study Site 1

	Sen	Spec	PPP	NPP	Total%	Kappa
Grade 3[a]						
Fall	-3.71*	3.24*	3.45*	-1.53	1.26	0.67
Winter	-3.94*	2.89*	3.23*	-1.82	0.38	-0.11
Spring	-3.31*	3.01*	2.80*	-1.37	1.04	0.40
Grade 4[b]						
Fall	-3.68*	2.17*	1.41	-1.94	-0.16	-0.49
Winter	-2.90*	1.44	1.00	-1.46	-0.46	-0.65
Spring	-1.55	1.86	1.95	-0.70	-0.49	0.28
Grade 5[c]						
Fall	0.00	0.00	0.00	0.00	0.00	0.00
Winter	1.07	-0.33	0.14	0.43	0.43	0.38
Spring	0.68	-0.82	-0.66	0.31	0.00	0.05

Note. Sen = sensitivity; Spec = specificity; PPP = positive predictive power; NPP = negative predictive power.
[a]n = 85. [b]n = 73, [c]n = 89.
*p < .05, two tailed.

The z-score test results at Study Site 2 are illustrated in Table 32. Within this data set, significant differences in

the total accuracy percentage were found between the locally-generated benchmarks and DIBELS benchmarks in fall, z = 2.39, *p* = 1.96, winter, z = 2.05, *p* = 1.96, and spring, z = 2.12, *p* = 1.96. When the prediction accuracy was corrected for chance agreement using kappa, significant differences were not identified for grade 3, z = .69 - .97, *p* = 1.96. Significant differences were not found in grade 4 between either the total accuracy percentages, z = 1.31 – 1.94, *p* = 1.96, or between the kappa values, z = −0.38 - 0.16, *p* = 1.96. Analysis of grade 5 data showed similar results as significant differences were not identified between the total accuracy percentages, z = −0.60 - .21, *p* = 1.96, or the kappa values, z = −1.66 - 0.04, *p* = 1.96. Therefore the null hypothesis was accepted.

Table 32

z-Score Tests for Differences between DIBELS- and Locally-Generated Benchmark Diagnostic Accuracy Statistics in Study Site 2

	Sen	Spec	PPP	NPP	Total%	Kappa
Grade 3[a]						
Fall	-4.79*	4.49*	8.08*	-1.06	2.39*	0.86
Winter	-2.98*	3.58*	7.58*	-0.68	2.05*	0.97
Spring	-4.79*	4.11*	6.41*	-1.35	2.12*	0.69
Grade 4[b]						
Fall	-6.58*	4.37*	6.13*	-1.33	1.70	-0.38
Winter	-2.03*	3.08*	6.58*	-0.24	1.31	0.16
Spring	-3.94*	4.29*	11.21*	-0.50	1.94	0.13
Grade 5[c]						
Fall	-7.18*	3.06*	1.00	-2.42*	-0.60	-1.66
Winter	-2.89*	1.76	1.19	-0.83	-0.20	-0.62
Spring	-0.48	0.91	0.88	0.00	0.21	0.04

Note. Sen = sensitivity; Spec = specificity; PPP = positive predictive power; NPP = negative predictive power.
[a]*n* = 129. [b]*n* = 122. [c]*n* = 127.
*$p < .05$, two tailed.

Differences between diagnostic accuracy statistics

Tables 30 and 31 also include the z-score tests between the diagnostic accuracy statistics calculated for the DIBELS and locally-generated benchmarks. Although significant differences were not present between the majority of the total accuracy percentages and kappa values for the local and DIBELS benchmarks, this pattern was not found in relation to the diagnostic accuracy statistics. In both study sites, significant differences were identified between the levels of sensitivity, specificity and PPP in each time period assessed.

In third grade in Study Site 1, The DIBELS benchmarks showed significantly greater levels of sensitivity, $z = -3.31 - (-3.94)$, $p = 1.96$, than what was produced by the DIBELS benchmarks. The reverse was true for specificity, $z = 2.98 - 3.24$, $p = 1.96$ and PPP, $z = 2.80 - 3.45$, $p = 1.96$. Although the NPP of DIBELS benchmarks outperformed the locally-generated benchmarks, significant differences were not obtained on this measure, $z = -1.37 - (-1.82)$, $p = 1.96$.

A similar pattern of performance was found in grade 4 in Study Site 1. At each time period assessed, the DIBELS benchmarks demonstrated higher levels of sensitivity and NPP while the locally-generated benchmarks produced higher levels of specificity and PPP. The differences, however, were not significant in all time periods across all statistics

calculated. Results of the z-score test for sensitivity yielded significant differences in the fall, z = −3.68, *p* = 1.96, and winter, z = −2.90, *p* = 1.96. A significant difference was identified for specificity only in the fall, z = 2.17, *p* = 1.96. No other significant differences were identified. As nearly identical benchmark scores were found in the grade 5, no significant differences were found in this grade level.

In Study Site 2, significant differences were found between the local and DIBELS benchmarks for sensitivity, specificity, and PPP. Similar to the findings of Study Site 1, the third grade DIBELS benchmarks significantly outperformed the locally-generated benchmarks in levels of sensitivity z = −2.98 – (−4.79), *p* = 1.96 at each time period assessed. The DIBELS benchmarks also showed significantly higher levels of NPP, although the differences were not significant, z = −0.68 – (−1.35), *p* = 1.96. The locally-generated benchmarks again produced significantly higher levels of specificity, z = 3.58 – 4.11, *p* = 1.96, and PPP, z = 6.41 – 8.08, *p* = 1.96.

This pattern was repeated in grade 4 in Study Site 2. The DIBELS benchmarks showed significantly higher levels of sensitivity, z = −2.03 – (−6.58), *p* = 1.96. The locally-generated benchmarks produced significantly higher levels of

specificity, z = 3.08 - 4.37, p = 1.96, and PPP, z = 6.13 – 11.21, p = 1.96. The DIBELS benchmarks showed higher levels of NPP in grade 4 but the differences as represented by the z-scores were not significant, z = –0.05 – (–1.33), p = 1.96).

The differences between the local and DIBELS benchmarks were less consistent in grade 5 in Study Site 2. In the fall, the DIBELS benchmarks produced significantly greater levels of sensitivity, z = –7.18, p = 1.96, and NPP, z = –2.42, p = 1.96. The locally-generated benchmark produced significantly higher levels specificity, z = 3.06, p = 1.96. Differences in PPP were not significant, z = 1.00, p = 1.96. In the winter, the DIBELS benchmarks generated significantly higher levels of sensitivity, z = –2.89, p = 1.96. Significant differences were not found between values for specificity, NPP, and PPP. No significant differences were found in the spring of grade 5 as the benchmark scores provided by the DIBELS system and the calculated locally-generated benchmark were nearly identical.

Summary

In this chapter, the analyses used to answer the research questions were discussed. A series of independent samples t-tests were conducted to assesses whether significant differences were present between the mean ORF scores at both study sites. Contrary to the hypothesis stated in Research Question 1, significant differences were found between the two

study sites with Study Site 2 consistently earning higher mean ORF scores. Given these results, analysis of the three remaining research questions was conducted for both study sites separately. Pearson correlations were computer to answer Research Question 2. The correlations between fall, winter, and spring ORF measures were very strong, greater than .90 in all cases. Correlations between ORF and the PSSA were moderate to strong. Confirming the stated hypothesis, these correlations were consistent with those found in the existing research literature with coefficients ranging from .64 - .70 in third grade, .54 - .64 in fourth grade, and .57 - .62 in fifth grade. In addition, the correlations demonstrated a similar diminishing pattern as found in previous research studies that show stronger correlations in grade 3 with weaker correlations in grades 4 and 5.

A logistic regression procedure was conducted to develop locally-generated benchmark scores. Consistent with the hypothesis stated in Research Question 3, the developed locally-generated benchmark scores were lower than those produced by the DIBELS system in grades 3 and 4. In grade 5, the locally-generated benchmarks were consistent or higher than the DIBELS benchmarks. Diagnostic accuracy statistics were calculated to analyze the balance of false positives and false negatives between the benchmark scores. Significant

differences between the local and DIBELS benchmark scores were calculated using z-score tests between the total accuracy percentages and values for kappa in both study sites. No significant differences were found between the locally-generated benchmarks and DIBELS benchmarks in Study Site 1. Significant differences were identified between the total accuracy percentages calculated for the locally-generated benchmarks and DIBELS benchmarks in Study Site 2 for grade 3 only. These differences were not sustained when the kappa values, which adjust for chance agreement, were analyzed. Statistically significant differences were present between diagnostic accuracy statistics in both study sites. A consistent pattern was identified showing that the levels of sensitivity and negative predictive power of the Grades 3 and 4 DIBELS-generated benchmarks generally outperformed those obtained from the locally-generated benchmarks. Conversely, the grades 3 and 4 locally-generated benchmarks showed higher levels of specificity and PPP than was produced by the DIBELS benchmarks. Significant differences were not found between the majority of the diagnostic accuracy statistics calculated for the locally-generated benchmarks and the DIBELS benchmarks in grade 5.

CHAPTER V

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Given the importance of the decisions made with the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) benchmarks, more information is needed to understand if the nationally-derived benchmarks created by the DIBELS system provide the most accurate criterion for evaluating reading proficiency, particularly when applied at the state or school district level. The DIBELS benchmarks are calculated based on performance with a nationally-normed standardized achievement test (Good et al. 2011b). Although this procedure may reflect reading proficiency at the national level, it may not accurately represent the standard of performance on a state assessment. Using a nationally-normed assessment like the DIBELS benchmarks may show a high number of false positives or false negatives when using the benchmark score to predict state test scores. Therefore, a different criterion that more accurately reflects local expectations may be needed. A benchmark expectation, established using a state assessment as a criterion for success, may provide a valid alternative to the nationally-normed benchmark created by the DIBELS system.

It is important to note that setting goals to meet a local state assessments and expectations is not without its limitations. This is because the state assessments and the

standard on which they are based may not be a valid benchmark to judge student achievement. In a 2009 Time magazine article, Walter Isaacson discussed this issue and provided an argument for how to raise the standards in America's school. He proposed that the No Child Left Behind (NCLB; 2001) legislation to raise academic standards through its push for universal proficiency may have resulted in unintended negative consequences. Although the rationale behind NCLB is admirable, the falling standards in American schools may have been worsened by the legislation. This is because each state has been permitted to individually decide how proficiency will be defined and assessed.

Isaacson's argument is strongly supported by Kingsbury et al. (2004). Kingsbury et al. found that a great disparity exists in the difficulty of the proficiency assessments from one state to the next. Although some states set high standards, in others proficiency on a state test is equivalent to a score at the 13th percentile on a nationally-normed test of achievement. Unfortunately, too many states have taken the latter approach and have set low state standards to satisfy proficiency requirements. This is because state assessments designed to meet these low expectations will show an inflated level of student achievement (Isaacson, 2009). The most troubling consequence of these low standards is that many

students are left without the basic skills necessary to compete in a global market even though state assessment scores suggest proficient performance. To help correct this trend, Isaacson suggests that a push toward a rigorous and valid national standards and assessments to which all students can be judged and measured. This would help to ensure that all students are meeting valid proficiency standards and no child is truly left behind in their academic development.

The PSSA, used in this study is a state assessment based on a set of standards that were selected and developed by the Pennsylvania Department of Education. By no means does the PSSA represent what might be considered the "gold standard" of reading assessments. It is merely an assessment that measures students' progress toward what the Commonwealth of Pennsylvania has determined is an adequate level of achievement. With this in mind, the PSSA was still chosen as the standard to which locally-generated ORF benchmarks would be developed in this study. This is because, whether or not the PSSA is a valid assessment of academic proficiency, it is the primary assessment to which students, teachers, and school districts are judged in Pennsylvania. It is therefore essential for school personnel to simultaneously understand the limitations of the PSSA but also to ensure progress is made by students toward this end of the year assessment.

The PSSA and ORF data used in this study were collected from two school districts in central Pennsylvania. Four research questions were proposed to determine if a local criterion for performance would be beneficial. The first two questions were put forward to understand the relationship between ORF and the PSSA. The purpose was to create a research-supported foundation on which locally-generated benchmark scores could be constructed. If a strong relationship is present, then the application of an ORF benchmark can be viewed as a valid standard to predict PSSA proficiency. To accomplish this goal, the ORF scores collected from both study sites were analyzed using independent samples t-tests to determine if significant differences were present. This analysis would help decide whether or not the scores from the two sites could be aggregated to create one set of benchmarks or if data from the sites should be analyzed separately. As significant differences were identified between the ORF scores at both study sites, the remaining research questions were answered by analyzing the data from each study site separately. The second research question was posed to determine the strength of the relationship between ORF and the PSSA. To answer this question, Pearson correlation coefficients were calculated

between the fall, winter, and spring ORF measures at both study sites in grades 3 -5.

With the foundation set, the third research question was posed to create the local benchmark expectations using state assessment performance as the criterion for success. To evaluate this question, student data from the two study sites were divided into two groups titled the Benchmark group and the Comparison group. This split was accomplished using random case selection in SPSS. The groups were stratified based on the students' level of performance on the PSSA. Once the groups were established, the data from the Benchmark groups were analyzed using a logistic regression procedure. Through logistic regression the scores that produced the highest overall accuracy percentages were identified for the fall, winter, and spring in grades 3-5 in both study sites. These selected scores represent the locally generated benchmarks. Once developed, the locally-generated benchmarks were applied to the Comparison group data. The accuracy of these locally-generated benchmark scores to predict the PSSA proficiency of the Comparison group students was then analyzed in Research Question 4. Diagnostic accuracy statistics, including sensitivity, specificity, NPP, positive predictive power, and total accuracy percentage were collected. In addition, values for kappa and phi were produced. The primary

goal of Research Question 4 was to understand the extent of the differences between the generated locally-generated benchmark scores and the benchmark scores provided by the DIBELS system to determine which set of benchmark scores best predicted PSSA proficiency. To accomplish this goal, z -score tests were calculated between the overall prediction accuracy percentages and the kappa values at both study sites. In addition, z-score tests measuring the differences between the diagnostic accuracy statistics were also calculated to provide a more detailed level of analysis.

The remainder of this chapter will discuss the analyses presented in Chapter 4. The results of each research question will be discussed in greater depth and will be integrated with the extant, relevant literature. An explanation of the findings will be provided as well as a discussion of the limitations, implications, and recommendation for future research will be included.

Research Question 1

Are there statistically significant mean differences on DIBELS benchmarks between the two participating schools? This first research question was proposed to determine if the data collected from the individual study sites could be aggregated into one dataset for subsequent analyses. If significant differences were identified between the ORF scores, then the

remaining research questions would be answered by analyzing the data from each study site separately. If significant differences were not identified, the data for both study sites would be aggregated and used to answer the remaining research question.

The stated hypothesis for this research question proposed that differences would not exist between the benchmark scores generated for the participating school districts. This hypothesis was made because the study sites share relatively similar racial, sex, and socio-economic demographic characteristics. In addition, analogous percentages of special education students and English Language Learners are present in both study sites.

To answer this research question, a series of independent samples t-tests were conducted. The t-tests were carried out between both study sites at each of the three benchmark assessment assessments and at each grade level. Contrary to the stated hypothesis, significant differences were identified between nearly all of the benchmark scores except for the fall and winter of fifth grade. During these two assessment periods, mean differences of only 5 words correct per minute (wcpm) were identified. Between the remaining benchmarks, mean differences ranged from 10 to 22 Study Site 2 consistently produced higher mean ORF scores than Study Site

1. This trend was also reflected in rates of PSSA proficiency where mean PSSA scores in Study Site 1 fell significantly below those found in Study Site 2.

The significant differences produced by the majority of the t-test lead to the rejection of the stated hypothesis for Research Question 1. It was believed that the demographic similarities between the two study sites would yield similar levels of performance. This pattern, however, was not found. The demographic similarities between the two districts did not correspond to similarities in reading achievement as measured by DIBELS or PSSA. Rather than analogous levels of performance, student ORF scores in Study Site 2 were significantly higher than those in Study Site 1 at all grade levels and at nearly every time period assessed. This same pattern was reflected in the percentage of students who had passed the PSSA. In Study Site 1, an average of 65% of students passed the PSSA. In Study Site 2, the average percentage of students who passed the PSSA was 81%.

Several possible causes could explain the disparity between levels of ORF at both study sites. These differences may include but are certainly not limited to, differences in the quality of the curriculum, availability of support personnel, access to high-quality instructional materials and differences in the quality of the instruction that is being

delivered. Opportunities for professional development, parental support and involvement, and unconsidered student characteristics may also be contributing to the differences between the study sites. Differences in socio-economic and environmental characteristics may also be contributing to the disparity between the two study sites. According to Paleologos and Brabham (2011) differences in income are strongly related to differences in reading achievement. In the present study, the median household income of Study Site 2 was nearly $24,000 per year higher than the income earned in Study Site 1 (Proximity, 2013). The vast difference in income earned within the two school districts may reflect discrepancies within the home learning environments such as differing levels of vocabulary exposure, familial support, and access to educational opportunities and materials (Paleologos & Brabham, 2011). These discrepancies may have manifested as differences in reading achievement. Further evaluation, beyond the scope of this study, may be necessary to fully examine the extent and causes of the disparity in reading achievement between the two school districts.

The primary goal of this research question was to determine whether or not the student data could be combined to create one data set that would be analyzed in the remaining research questions. The differences between the two study

sites provided sufficient evidence to suggest that the data lacked the consistency necessary to combine the study sites into one large data set. Consequently, the data collected for both study sites was analyzed separately for the remaining research questions. Therefore, the correlations, the benchmark scores, and the analyses that followed, were interpreted at the school district level. It is believed that this separation created a more individualized level of analysis. This level of analysis may be easier to interpret by school personnel as it reflects students within their own school building who experience the same instructional environments. Unfortunately, the differences between the school districts may have one important negative consequence. Separating the data from the study sites resulted in the sample sizes used to generate the benchmark scores were reduced by 50%. This was particularly important as the data set was already planned to be divided in half for the purpose of cross validation. This especially affected the data collected in Study Site 1 which had fewer students within each grade level. This smaller sample size may have skewed their results.

The significant differences between the data collected for both study sites, however, has another important consequence. Because of the disparity between the ORF scores

a stronger argument can be made to apply a local benchmark expectation rather than just employing a national benchmark. This supports Silberglitt's (2008) contention that school districts should avoid simply adopting a national benchmark standard. This is because the one-size-fits-all national benchmark score will not necessarily correspond to local expectations of student performances in all settings with all students equally. As exemplified in this research question, student skills can from one school district to another can vary greatly. One benchmark will likely not be sufficient to accurately judge student performances in both higher and lower performing school districts.

Research Question 2

What are the correlations between the fall, winter, and spring DIBELS ORF scores and performance on the PSSA in grades 3 - 5? Consistent with previous research (Baker et al., 2008; Deno et al., 1982; Hintz & Silberglitt, 2005; Hosp & Fuchs, 2005; Jenkins & Jewell, 1993; Shinn et al., 1992), it was hypothesized that ORF at each grade level would demonstrate moderate to strong correlations with the results of the PSSA.

Analysis of this research question confirmed the stated hypothesis. Correlations between the fall, winter, and spring ORF measures were strong, generally above .90. The correlations were strongest in third grade. This pattern was

supported by the results of correlation analyses at both study sites. Consistent with the extant literature, results from this study further indicate stronger correlations between DIBELS and high-stakes tests in third grade with diminishing correlations in subsequent years (Jenkins & Jewell, 1993).

Jenkins and Jewell (1993) suggested possible reasons for this phenomenon. First, ORF is a more sensitive measure in the early grades than it is in the later elementary grades. Reduced variance in the scores would likely cause a decrease in the relationship between ORF and general reading achievement. Second, differences may be present in the capability of both types of assessment to measure growth in reading. ORF may not have a high enough ceiling to adequately assess student reading growth in the upper grades. At this point, increases in wcpm no longer lead to corresponding increases in general reading achievement. Third, reading achievement likely reflects different and more complex reading skills in the upper elementary grades. In grade 3, general reading proficiency assessments may emphasize vocabulary and word recognition skills that are not assessed in the later grades. In the primary grades, students are still acquiring the basic skills necessary for successful reading. Reading assessments, therefore, are mainly focused on measuring students' acquisition of those skills. These types of reading

tasks likely have a strong relationship with ORF fluency. Consequently, when students are able to read fluently, they are also able to efficiently apply phonemic awareness and decoding skills and access their reading vocabulary successfully. In the later elementary grades, students are now expected to apply these basic skills to fluently read text and to comprehend what they are reading. Reading assessments in these grade levels begin to focus less on basic skills development (i.e., decoding, fluency) and more on comprehension skills that are not adequately represented by ORF alone.

Research Question 3

What are the locally-generated benchmark scores in the fall, winter, and spring in grades 3 - 5? Logistic regression was used to calculate the locally-generated benchmarks. This procedure was chosen because the cut score in a range of student ORF scores that produces the highest percentage of correct predictions of the PSSA is selected and utilized as part of the analysis with logistic regression.

It was hypothesized that the locally-generated benchmarks would be lower than those created by the DIBELS system. This is because the DIBELS benchmarks are designed as a universal screening tool that is purposely inflated to ensure a higher number of students will receive intervention

supports (Good et al., 2011b). This inflated cut score, useful for screening purposes, sacrifices the accuracy of the prediction of PSSA proficiency (Hintze & Silberglitt, 2005). In addition, the logistic regression used in this study maximized the percentage of true positives only and produced a benchmark score that was not artificially inflated but maximized the prediction accuracy on the PSSA. According to Hintze and Silberglitt (2005), logistic regression produces cut scores which are lower than other methods of calculation.

Study Site 1.

The grades 3 and 4 locally-generated benchmarks were as many as 36 wcpm lower than those produced by the DIBELS system. Lower benchmarks than those generated by either Study Site 1 or the DIBELS system were found in Study Site 2. This was not unexpected given that the student performances in Study Site 2 were consistently greater than their Study Site 1 counterparts. The locally-generated benchmarks in both study sites were not substantially lower than the DIBELS benchmarks as the differences between the scores were primarily not greater than one standard deviation.

The differences between the locally-generated benchmarks and the DIBELS benchmarks in Grade 5 were slight. This was particularly evident in Study Site 1 were the locally-generated benchmarks equaled or exceeded those of the DIBELS

System. In Study Site 2, the differences between the locally-generated benchmark and the DIBELS benchmarks diminished from the beginning to the end of the year. When the students were assessed in the spring with ORF, almost no difference was found between the locally-generated benchmark and the DIBELS benchmark

The decreasing differences between the DIBELS benchmark and locally-generated benchmark are due in large part to the diminishing relationship between ORF and general reading achievement. In grade 3, the locally-generated benchmark is lower than the DIBELS benchmark because reading achievement at this grade level is heavily influenced by basic reading skills like phonemic awareness, understanding of the alphabetic principal, and vocabulary knowledge (Jenkins & Jewell, 1993). Successful students in grade 3 are those who are able to efficiently harmonize and apply these skills when reading. ORF is particularly adept at measuring this efficient application because to read quickly and efficiently one must have mastered each of these basic skills. In subsequent grades, the relationship between ORF and overall reading achievement diminishes for the reasons discussed in the previous research question. ORF, as a solitary indicator, is no longer sufficient to represent the myriad abilities and skills needed for successful reading in the later grades.

Mastery of the basic skills becomes a foregone conclusion and reading achievement is now measured by the student's proficiency in comprehending what he /she has read. Consequently, the ability to apply an ORF benchmark expectation to predict general reading achievement also diminishes. Furthermore, the locally-generated benchmark scores used were selected because the logistic regression procedure used to create the benchmarks identified the scores that maximized the percentage of true positives. As the relationship between ORF and the PSSA weakens, the benchmark level, as calculated through logistic regression, must rise to compensate for this declining relationship. The number of words a student is able to read in one minute may continue to increase; however, increases in ORF do not necessarily translate into proficiency on the PSSA. Many students who are able to read with a higher level of ORF will still fail the PSSA. To ensure a high number of true positives, the benchmark score, calculated through logistic regression, must increase to account for the higher number of students who show higher levels of ORF but who still fail to meet proficiency on the PSSA.

Research Question 4

Are the locally-generated benchmarks able to predict PSSA proficiency with significantly greater accuracy than the

DIBELS benchmarks? Additionally, are measures of diagnostic accuracy (sensitivity, specificity, PPP, and NPP) significantly different based on the derivation of the benchmarks? It was hypothesized that significant differences will be identified between the locally-developed benchmarks and the DIBELS-generated benchmarks in their ability to reliably predict PSSA performance. It was further hypothesized that the locally-generated benchmarks would more accurately predict PSSA performance. In addition, significant differences will be present between the diagnostic accuracy statistics for both sets of benchmarks. These hypotheses were suggested for two reasons. First, the DIBELS benchmarks represent a set of standards that are designed to ensure that struggling students receive necessary interventions (Good et al., 2011b). To meet this goal, the developers of DIBELS inflate their benchmark expectations to make sure that a higher number of students will be identified as at-risk and be given access to these interventions. Although appropriate for screening decision, the inflation of the DIBELS benchmark score decreases the accuracy of the prediction of PSSA performance. Second, locally-generated benchmarks developed by Ferchalk et al. (2010) more accurately predicted proficiency on the PSSA than DIBELS-generated benchmarks. Similar findings were predicted for this study.

Overall Prediction Accuracy

The stated hypothesis suggested that the locally-generated benchmarks would more accurately predict PSSA performance than those created by the DIBELS system. The results of the z-score tests using data from both study sites indicated that this hypothesis was not supported. In Study Site 2 significant differences were identified between the total accuracy percentages of the locally-generated benchmarks and the DIBELS benchmarks in grade 3. The differences, however, were not maintained in subsequent grades. In addition, when total accuracy percentage was adjusted for chance using kappa, no significant differences were identified. No significant differences were identified between the total accuracy percentages or kappa values of the locally-generated benchmarks and DIBELS benchmarks in Study Site 1. In grade 5, the difference between the DIBELS and locally-generated benchmarks were greatly reduced. There was no discernible difference between the local and DIBELS benchmarks in Study Site 1. In some cases the locally-generated benchmark was higher than the DIBELS benchmarks. In Study Site 2, the locally-generated benchmark remained lower than the DIBELS benchmark though not significantly lower.

The reason that no differences were found between the preponderance of the benchmarks is likely due to the

fluctuation between false positives and false negatives. The logistic regression procedure used to calculate the benchmarks scores produced a cut score that maximized the percentage of students who were predicted to fail the PSSA and actually failed the PSSA (true positives). One of the side effects of amplifying true positives is a corresponding increase in the amount of false negatives. When the threshold is lowered the established benchmark score will lessen the possibility of students who will be predicted to pass the PSSA but who subsequently fail. The reverse is true when the benchmarks scores are raised. By raising the threshold, the number of students predicted to pass the PSSA based on their ORF score but who actually fail is reduced. The consequence of this precarious balancing act is that the total prediction accuracy of the local and DIBELS benchmarks are similar and any differences between the two are generally not statistically significant. Rather than identify a benchmark score that maximizes overall prediction accuracy, the raising or lowering of the benchmark tips the balance of false positives and false negatives without significantly altering overall prediction accuracy.

Differences in Sensitivity and Specificity

Given that significant differences were not present between total prediction accuracy of the benchmark scores a

second level of analysis was conducted examining the differences between sensitivity and specificity. The z-score tests yielded significant differences between the levels of sensitivity and specificity calculated for both the locally-generated benchmarks and DIBELS benchmarks. This is unsurprising given the differences in where the benchmarks expectations were drawn. In grade 3 the difference between the two benchmarks in Study Site 1 was an average of 22 wcpm with a 24 wcpm difference in Study Site 2. In both cases, the locally-generated benchmarks were lower than the DIBEL-generated benchmarks. In grade 4, the difference between the two benchmarks was an average of 14 wcpm in Study Site 1 and 33 in Study site 2. As discussed in the previous section, by decreasing the benchmark, the balance between sensitivity and specificity is altered. A lower benchmark increases specificity because it maximizes the accuracy of true positive predictions. This indicates that if a student falls below the lower ORF benchmark one can be very confident that the student will not perform successfully on the PSSA. Unfortunately, the reverse is also true when the benchmark scores are lowered. The ability of the benchmark to accurately identify students who will pass the PSSA is affected. The lower score will produce a threshold for performance that ensures more students will be identified as low risk. This will increase the number

of false negatives as a higher number of students will fail the PSSA even after meeting the benchmark expectation.

Conversely, the higher DIBELS benchmark increases sensitivity. This higher cut score produces more accurate true negative predictions. Using this higher benchmark score ensures a different level of interpretation. One can be more confident that a student who is able to perform above the higher DIBELS ORF benchmark will meet proficiency on the PSSA. This threshold is more exclusive for successful readers ensuring only those who are very fluent will be predicted to pass the PSSA. This leads to a lower percentage of false negatives and will simultaneously ensure that a greater number of students will be identified as at-risk for reading failure. Many of these at-risk students, however, will still pass the PSSA. Therefore, raising the benchmark to increase true negatives will simultaneously increase the number of false positives.

Through this analysis it is clear that neither the DIBELS nor the locally-generated benchmarks are able to satisfactorily amplify overall accuracy percentage while simultaneously maximizing both sensitivity and specificity. This is particularly true in grades 3 and 4 where significant differences were identified between the levels of sensitivity and specificity. Depending on where the benchmark is drawn,

the number of false positives or false negatives will increase. Although this pattern is a weakness inherent in using benchmark cut scores, it can be re-framed as a strength when employed and interpreted properly. Therefore, rather than simply employing a one-size-fits-all benchmark, school personnel should consider using more than one.

The choice between locally-generated benchmarks and DIBELS benchmarks will depend on how the data are to be employed. Benchmark scores are primarily used for three different purposes: screening decisions to determine which students are in need of supplemental intervention supports, high-stakes decision-making such as special education determinations, and administrative decision-making purposes that evaluate the quality of the curriculum provided to the students. Unfortunately applying one benchmark to meet each of these three purposes may not be feasible. One benchmark may effectively meet one or two of these purposes but not all three. The higher DIBELS benchmark will be more appropriate for use as a universal screening tool. This is because the higher threshold ensures that more students will be identified in need of additional reading supports. It is important to note that it is likely that some of the students who fall below a higher benchmark would pass the PSSA even if additional reading supports were not provided. It is

difficult, however, to identify these false positives before the PSSA is taken. It is therefore preferable to provide an intervention to a student who may not need it rather than to overlook a student who will fail if supports are not provided.

The lower locally-generated benchmark has the potential for a separate use. Because of the higher levels of specificity, the accuracy of true positive predictions made by the locally-generated benchmark is maximized. This ensures that students who fall below the benchmark have very low probability of passing the PSSA. This information is particularly valuable for special education determinations within a Response to Intervention model (RtI).

RtI refers to a school improvement paradigm that employs a multi-tiered service delivery model used to employ a system of research-based core and supplemental interventions to meet the needs of students (Tilly, 2006). Within this model, RtI frequently employs curriculum-based measurement (CBM) to determine which students are in need of supplementary interventions and to monitor student progress after the interventions are implemented (Daley, Martens, Barnett, Witt, & Olsen, 2007; Lichtenstein, 2008). Students, who do not show adequate response to the intervention as measured by CBM, may be eligible for special education supports as a student with a specific learning disability (SLD).

To determine SLD eligibility within an RtI model, Fuchs and Fuchs (1998) recommend a dual-discrepancy approach. This model evaluates two important features of achievement: level of performance compared with same aged peers and rate of learning (Fuchs, 2003; Fuchs & Fuchs, 2007; Speech, 2003; Speech & Case, 2001). As part of a SLD evaluation, the determination of insufficient rate of improvement is calculated by measuring student progress on reliable measures over a sufficient period of time (Fuchs & Fuchs, 2007; Fuchs, Fuchs, Hamlett, Waltz, & Germann, 1993). To evaluate an insufficient level of performance, the evaluator may consider the student's CBM score in relation to a pre-selected benchmark criterion (Burns, 2008). The DIBELS benchmark, however, may be inappropriate for this use as it may be an unfairly inflated score given the number false positives produced when used to predict PSSA proficiency. On the other hand, the lower locally-generated benchmark significantly reduces the number of false positives made when predicting PSSA proficiency. Under these circumstances, the locally-generated benchmark is preferable as a determinant for an insufficient level of performance because the evaluator can be sure that if a student falls below this threshold of performance, he / she will likely fail to meet proficiency.

As school personnel interpret the benchmark scores, they may determine that a student who falls below the DIBELS benchmark is an unsuccessful reader. This incomplete level analysis, however, can be inaccurate and too simplistic. Administrators should consider these locally-generated benchmarks to give a better understanding of student performances within their school district. As opposed to the DIBELS benchmark, a student who falls below the locally-generated benchmark is very unlikely to pass the PSSA. This information can be very valuable to help determine the success or failure of instructional programs. Conversely, the DIBELS benchmark is likely a better threshold to judge which students are likely to pass the PSSA. The locally-generated benchmark produces far too many false negatives to be useful for this purpose. The DIBELS benchmark, on the other hand, produces a higher percentage of true positives. This helps to ensure that if a student is able to exceed this benchmark scores, than they will be highly likely to pass the PSSA.

Fuchs, Fuchs, and Compton (2012) discussed the potential for what they define as "Smart RtI." Rather than a one-level approach to universal screening in RtI, they propose the use of a multi-stage method to screening to more accurately determine the students most in need of additional supports. They contend that current screen practices using CBM have a

great potential for a high number of false positives when employing a benchmark criterion. This high level of error may force school districts to purchase costly interventions and use already scarce resources to intervene with students who do not need the supports to be successful. Fuchs et al. believe that employing a secondary level of screening assessments will reduce the number of false positives made through CBM benchmarking assessments. The result will help to ensure that only students who were most in need would receive supplemental supports, significantly reducing the financial burden on the school district.

The findings of this study support this contention posed by Fuchs et al. (2012). The application of a second level of assessment may indeed provide sound information that leads to more efficient RtI practices. Additional screening assessments, however, may not be necessary to meet this goal. More in-depth and efficient analysis of the CBM data currently collected may provide a similar "smart" approach to RtI as proposed by Fuchs and Fuchs. The application of two separate benchmark cut scores, as proposed in this study, would help to further delineate between those students who are most in need of intervention supports from those who are not. A process could be applied that first selects the students who fall below the DIBELS benchmark. This would provide a general

understanding of those students who are at-risk for reading failure. To further delineate between students who are identified as at-risk, the locally-generated benchmark score could be applied to more accurately select the students most in need of interventions. The low number of false positives produced by the locally-generated benchmark will ensure that only those students who are truly struggling in reading are identified as at-risk. This is particularly important for school districts with limited resources. Rather than spend valuable resources on those who are not in need, teachers will be able to more efficiently allocate their resources to the students who they can be sure will not meet proficiency without the help.

Limitations

Participants in this study included students in grades 3 - 5 who took the PSSA. Students who were assessed using modified version of the PSSA or the Pennsylvania's Alternative System of Assessment (PASA) were not included in this study. Their exclusion may have skewed the results of this study as the lowest performing students with learning disabilities or intellectual disabilities were not included in the analysis. The potential uses of the present study may not be applicable to these populations.

Approximately 85% of students included in this study were identified as White/Non-Hispanic. The remaining 15% were comprised of students from other racial backgrounds including Asian/Pacific Islander, Hispanic, and Black/Non-Hispanic students. Although students from these racial backgrounds were included, their minimal representation may not be sufficient to suggest that the results of this study are generalizable to students from these populations. Extant literature shows that an achievement gap, though steadily declining due to improvements in education over the past 20 years, still exists between White students and their African-American and Hispanic counterparts (Harris & Herrington, 2006; J. Lee, 2002). These differences may affect the level of performance present with ORF scores and state test performance and the connection between the two measures. Therefore, replications are recommended in settings that represent a higher representation of racial and ethnic minorities to ensure consistent findings.

No data were collected during this study regarding the accuracy of the assessment administration procedures. This includes the administration procedures of both the PSSA and the DIBELS assessments. Therefore, no evidence was presented in this study that validated the accuracy of the data

collected. Caution must therefore be taken when interpreting results obtained from these data.

Large differences were present between the ORF and PSSA data collected for both study sites. These differences precluded aggregation of the data across study sites. This disaggregation of data across study sites reduced the overall sample size by a substantial degree and limited the size of the cross-validation sample used to develop the locally-generated benchmarks. Similarly, given the smaller sample size, student scores were dichotomized into pass/fail categories rather than reflect the four performance levels of the PSSA. This analysis may limit the usefulness of the information provided as it does not delineate between students who performed in the Basic and Below Basic levels or between the students who performed in the Advanced and Proficient levels.

In addition, Study Site 2 was a very high achieving District with 75 – 85% of students scoring Proficient or Advanced on the PSSA. Though admirable, their performances produced a weak distribution of scores in the lower ranges of reading achievement. Less than 10% of students performed in the Below Basic range on the PSSA with less than 20% of students in the Basic range. Given the lack of variability in

these performance levels, the locally-generated benchmark scores generated from these scores may have been skewed.

As discussed within this chapter, the relationship between ORF and scores on the PSSA are strong in grade 3 but weaken in subsequent grades. Consequently, the ability to apply an ORF benchmark to predict PSSA performance is also weakened. This is particularly salient in grade 5 where only negligible differences are present between the DIBELS benchmarks and locally-generated benchmarks. Caution is recommended when interpreting ORF benchmarks in these grades as ORF may not be sensitive enough to measure reading growth in the latter elementary grades. Additionally, the relationship between ORF and general reading achievement is not a perfect correlation. Therefore, error will always be present when extrapolating a prediction from a CBM to a performance on a state assessment. Caution is again recommended when interpreting these scores.

Implications for Research

Several directions for future research are recommended as a result of this study. First, replications are suggested to examine whether the findings of this study remain consistent in other settings and with other populations. Similar studies with larger populations may help to ensure that wider distributions of performance levels are present.

As described by Isaacson (2009) and as demonstrated by the findings of Kingsbury et al. (2003) state assessments have limitations which call into question their validity. They are designed based on standards that are selected by each individual state without the context of a national standard or national curriculum. The end product of 50 unique sets of academic standards is great disparities between how proficiency is defined and measured from one state to another. These disparities weaken the use of a state assessment as a reliable and valid measure of student achievement. With this in mind, it is important to note that whether or not a state assessment is a valid measure of academic proficiency it is required by NCLB (2001) and remains the primary assessment to which students, teachers, and school districts are judged. It is therefore essential for school personnel to simultaneously understand the limitations of their state assessment and to ensure student progress is made toward this end of the year measure. Therefore, replications in other states are recommended to determine if the findings of this study are found in other states that use different state assessments. This will help to determine how well ORF benchmarks fit with local expectations across the country.

Studies with larger populations are also recommended for two important reasons. First, higher populations will help to

ensure that an equal distribution of student achievement is present. Both high achieving schools and low achieving schools should be included so that students from all skill levels are represented. Also, studies with greater sample sizes may consider analyses that do not dichotomize the student scores as pass/fail. This would allow for a more specified examination that will determine the usefulness of CBM benchmarks for each of the 4 PSSA performance levels: Below Basic, Basic, Proficient and Advanced.

Given the exclusion of lower-performing students who were assessed with the PASA and the modified version of the PSSA future research studies should evaluate benchmark expectations with this population of students. In addition, studies that account for the accuracy of the assessment administration should be conducted to ensure that the measures are administered according to standardized procedures.

Ongoing research studies should evaluate the connection between ORF and reading achievement in the upper elementary grades. As the relationship between ORF and reading achievement weakens in these grades, other assessments that supplement or replace ORF as a primary general outcome measure should be considered. Maze reading assessments provide one possible option for this role. Several researchers have found that maze assessments can be used as a reliable and valid

measure of reading skills in the upper elementary grades (Jenkins & Jewell, 1993; Silberglitt et al., 2006; Wiley & Deno, 2005). Research conducted by Jenkins and Jewell (1993) supported this notion as they found that maze assessments are a more sensitive measure at higher grades than at lower grades levels. The Dynamic Measurement Group (DMG) who develops the DIBELS system have also sought to improve reading measurement in the upper grades. In their most recent edition of DIBELS, the DMG has added a maze comprehension measure called DIBELS Daze (Good et al., 2011). Additionally, they have unveiled the DIBELS composite score as a way of adding to the validity of their indicators. In the upper elementary grades, the composite score is a combination of ORF wcpm, reading accuracy, reading retell, and DIBELS Daze comprehension. This collective of scores may help to compensate for the weaknesses of each individual indicator alone. More research, however, is needed to ensure the validity of both the composite score as well as DIBELS Daze.

Future research should analyze what other local-level data could be added to CBM to effectively add to the strength of the prediction to PSSA proficiency. This local data could include teacher recommendations, classroom attendance, rate of homework completion, student self-reports, classroom reading grades, classroom core and content area grades and other

information. The addition of one or more of these subjective factors may improve the accuracy of proficiency predictions than one CBM could do on its own.

Additional research should also follow the longitudinal progress of the students included in this study. In particular, the students who show a consistent pattern of false positives or false negatives in subsequent years should be further evaluated. This information will permit an additional level of analysis to help determine the students who are in need of additional supports. Students who show consistent false negatives are those who are able to read fluently and thus pass the ORF screening measure, but will eventually fail to meet proficiency on the state assessment. The identification of these students may lead to monitoring practices that rely on measures other than reading fluency. This will help to make sure that these students are not overlooked and are given access to the supports that they may need. Conversely, students who show a consistent pattern of false positives are those students who fail to reach the benchmark in ORF but who still meet proficiency on the state assessment. Further analysis of these students would be invaluable to understand what skills they possess that allow them to successfully overcome their reading fluency deficits to perform successfully in overall reading achievement.

Implications for the Practice of School Psychology

Silberglitt (2008) states that school districts should "refrain from simply adopting a set of national target scores, as these target scores may not be relevant to the high-stakes outcomes for which their students must be adequately prepared" (p. 1871). The results of this study help to confirm the validity of this statement. The applications of DIBELS ORF benchmarks are wide. They are used for a variety of purposes including, universal screening, progress monitoring, program evaluations, and special education determination. Unfortunately, their validity in each of these applications is also varied. Their value is dependent on the uses to which they are applied. The DIBELS benchmarks in grades 3 and 4 produce a high number of false positives when used to predict PSSA scores. Consequently, their application as a part of special education determination may be limited as evaluators will need to ensure that these decisions are made with the most accurate information. This does not suggest that the DIBELS benchmarks are without value. Because of their higher standard, they may be appropriately applied for universal screening purposes. Their higher threshold ensures that more students receive supplementary services and fewer students are left without. Additionally, the DIBELS benchmarks may be more appropriately applied to answer questions about which students

are most likely to pass a proficiency examination. Their higher standard ensures only students who exceed the standard are likely to pass a proficiency examination.

The DIBELS benchmarks, however, are far less reliable when answering questions about the students who are most likely to fail the PSSA. Locally-generated benchmarks are a more appropriate standard for this purpose as they produced far fewer false positives than the DIBELS benchmarks. This is also what makes the locally-generated benchmark a better alternative for use in making special education eligibility decisions. Because of the high percentage of true positive predictions, the evaluator can be better assured that a student who falls below the locally-generated benchmark is highly unlikely to pass a state assessment.

Whether or not school psychologists choose to calculate and employ local CBM benchmarks in their districts, they should, at minimum, appreciate how benchmarking works. They should understand how the benchmarks are created and their most appropriate applications. Most importantly, school psychologists should learn how well the nationally-derived DIBELS benchmark corresponds with the local expectation held by their own state or school district. With this knowledge, school psychologists can guide their colleagues to interpret the CBM benchmarks appropriately. This will help to ensure

that sound decision practices are used to determine how to best meet student needs.

Summary

Four research questions proposed to investigative the relationship between ORF and the PSSA were further examined in this chapter. The results for each of the four research questions were presented along with a framework for their interpretation. The first two questions were put forward to understand the relationship between ORF and the PSSA. The purpose was to create a research-supported foundation on which locally-generated benchmark scores could be constructed. If a strong relationship was present, then the application of an ORF benchmark can be viewed as a valid standard to predict PSSA proficiency. The results of these research questions highlight two important points. First, differences in the level of achievement in both participating schools provided a justification for the division of the data into two separate study sites rather than one combined cohort. In addition, the results determined that subsequent research questions would be answered using data from each study site separately. Second, a strong relationship was present between ORF and the PSSA. The strength of the relationship, however, was at its peak in third grade and diminished in grades 4 and 5. This suggested

that the validity of an ORF benchmark score in grade 4, but particularly in grade 5, may be questionable.

With the foundation set, the third research question was posed to determine the level of performance on DIBELS ORF that would predict PSSA proficiency with the highest overall accuracy percentage. These calculated benchmarks were calculated using a logistic regression procedure. Through logistic regression the scores that produced the highest overall accuracy percentages were identified for the fall, winter, and spring in grades 3-5 in both study sites. The calculated benchmark scores were lower than those produced by the DIBELS system in grades 3 and 4. The diminishing relationship between ORF and the PSSA lead to congruent benchmark scores in grade 5 in both study sites.

Once developed, the locally-generated benchmarks were applied to a cross validation sample of students. The accuracy of these locally-generated benchmark scores to predict the PSSA proficiency of the Comparison group students was then analyzed. Diagnostic accuracy statistics, including sensitivity, specificity, negative predictive power, positive predictive power, and total accuracy percentage were collected. In addition, values for kappa and phi were produced. The primary goal of Research Question 4 was to understand the extent of the differences between the generated

locally-generated benchmark scores and the benchmark scores provided by the DIBELS system to determine which set of benchmark scores best predicted PSSA proficiency. Significant differences were not present between the overall accuracy percentages or kappa values produced by the majority of the local and DIBELS benchmarks. The absence of differences was likely due to the declining relationship between ORF and general reading achievement. Differences were identified between values for sensitivity and specificity at both study sites. The application of both sets of benchmarks was recommended as a result of these differences. The specific use for a locally-generated benchmark or a DIBELS benchmark was suggested to be dependent upon the purpose for which it is to be used or the question that is to be answered.

Upon completion of these analyses the limitations of this study were discussed. These limitations included homogeneous populations, exclusion of the lowest performing students, large differences between study sites, no confirmation of accurate assessment procedures, and the declining relationship between ORF and the PSSA. Each of these limitations was subsequently addressed as recommendations for future research. Additionally, recommendations for the application of this study to the practice of school psychology were presented. These

recommendations suggest that school psychologists should understand how benchmarks are designed and how well they conform to local expectations. This will ensure that any benchmark used to measure student growth will be applied and interpreted appropriately.

References

Alonzo, J., Tindal, G., Ulmer, K., & Glasgow, A. (2006). easyCBM online progress monitoring assessment system [Online software]. Retrieved from http://www.easycbm.com

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, *10*(18), 1-74. Retrieved from http://epaa.asu.edu/epaa/v10n18/

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York, NY: Macmillan.

Ardoin, S. P., & Christ, T. J. (2008). Evaluating curriculum-based measurement slope estimates using data from triannual universal screenings. *School Psychology Review*, *37*, 109-125.

Atkins, T. L., & Cummings, K. D. (2011). Utility of oral reading and retell fluency in predicting proficiency on the Montana Comprehensive Assessment System. *Rural Special Education Quarterly*, *30*(2), 3-12.

Baker, S. K., & Good, R. (1995). Curriculum-based measurement of English reading with bilingual Hispanic students: A validation study with second-grade students. *School Psychology Review 24*, 561-578.

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame'enui, E. J., & Beck, C. (2008). Reading fluency

as a predictor of reading proficiency in low-performing, high-poverty schools. *School Psychology Review*, *37*, 18-37.

Berry, R. (2008). *Assessment for learning.* Aberdeen, Hong Kong: Hong Kong University Press.

Braden, J. P., & Tayrose, M. P. (2008). Best practices in educational accountability: High-stakes testing and educational reform. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 575-588). Bethesda, MD: National Association of School Psychologists.

Bradley, R., Danielson, L. C., & Hallahan, D. P. (2002). *Identification of learning disabilities: Research to practice*. Mahwah, NJ: Lawrence Erlbaum.

Breakwell, G. M. (2006*). Research methods in psychology* (3rd ed.). Thousand Oaks, CA: Sage.

Buck, J., & Torgesen, J. (2003). *The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test* (FCRR Tech. Report No. 1). Retrieved from Florida Center for Reading Research website: http://www.fcrr.org/TechnicalReports/TechnicalReport1.pdf

Burns, M. K. (2008, October). *Data-based problem analysis and interventions within RTI: Isn't that what school psychology is all about?* Paper presented at the

Association of School Psychologists of Pennsylvania Annual Conference, State College, PA.

Burns, M. K., & Senesac, B. V. (2005). Comparison of dual discrepancy criteria to assess response to intervention. *Journal of School Psychology*, *43*, 393-406.

Burns, M. K., VanDerHayden, A. M., & Boice, C. H. (2008). Best practices in delivery of intensive academic interventions. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 1151-1180). Bethesda, MD: National Association of School Psychologists.

Burns, M. K., Wiley, H. I., & Viglietta, E. (2008). Best practices in implementing effective problem solving teams. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 1633-1644). Bethesda, MD: National Association of School Psychologists.

Casey, A., & Howe, K. (2002). Best practices in early literacy skills. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 721-735). Bethesda, MD: National Association of School Psychologists.

Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment*, 7, 303-323.

Daly, E. J., Martens, B. K., Barnett, D., Witt, J. C., & Olsen, S. C. (2007). Varying intervention delivery in response to intervention: Confronting and resolving challenges with measurement, instruction, and intensity. *School Psychology Review 36*, 562-581.

Das, J. P., Naglieri, J. A., & Kirby, J. R. (1994). *Assessment of cognitive processes*. Needham Heights, MA: Allyn & Bacon.

Data Recognition Corporation. (2011). *Technical report for the 2011 Pennsylvania System of School Assessment: Assessment handbook*. Harrisburg, PA: Data Recognition Corporation. Retrieved from Pennsylvania Department of Education website: http://www.portal.state.pa.us/portal/server.pt/community/technical_analysis/7447

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

Deno, S. L. (2003). Developments in curriculum-based measurement. *Journal of Special Education*, *37*, 137-161.

Deno, S. L., Mirkin, P., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, *49*, 36-45.

Dominguez de Ramirez, R., & Shapiro, E. S. (2006). Curriculum-based measurement and the evaluation of reading skills of

Spanish-speaking English language learners in bilingual education classrooms. *School Psychology Review 35*, 356-369.

Dynamic Measurement Group. (2010). *DIBELS Next benchmark goals and composite score*. Retrieved from http://dibels.org/papers/DIBELSNextBenchmarkGoals.pdf

Ellis, A. K. (2001). *Research on educational innovations* (3rd ed.). Larchmont, NY: Eye on Education.

Ehri, L. C. (2004). Teaching phonemic awareness and phonics: An explanation of the National Reading Panel meta-analyses. In P. McCardle & V. Chhabra (Eds.), *The voice of evidence in reading research* (pp. 153-186). Baltimore, MD: Paul H. Brookes.

Engelmann, S., Meyer, L., Carnine, L., Becker, W., Eisele, J., & Johnson, G.(1999). *Corrective reading program.* Columbus, OH: SRA/Mc-Graw-Hill.

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole.

Ferchalk, M. R., Cogan-Ferchalk, J. R., & Richardson, F., (2012, Fall). DIBELS Next and performance on high-states assessments. *ASPP InSight, 33*(1), 4-6.

Ferchalk, M. R., Richardson, F., & Cogan-Ferchalk, J. R. (2010, October). *Using oral reading fluency data to create an accurate prediction model for PSSA performance.*

Poster session presented at the fall conference of the Association of School Psychologists of Pennsylvania, State College, PA.

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.

Fletcher, J. M., Coulter, W. A., Reschly, D. J., & Vaughn, S. (2005). Alternative approaches to the definition and identification of learning disabilities: Some questions and answers. *Annals of Dyslexia, 54,* 304-331.

Fuchs, D. Fuchs, L. S., & Compton, D. L. (2012). Smart RTI: A next-generation approach to multilevel prevention. *Exceptional Children, 78*, 263-279. doi:

Fuchs, L. S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities Research and Practice, 18*, 172-186.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, *53*, 199-208.

Fuchs, L. S., & Fuchs, D. (1997). Use of curriculum-based measurement in identifying students with disabilities. *Focus On Exceptional Children*, *30*(3), 1-16.

Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying construct of reconceptualizing the identification of learning disabilities. *Learning Disability Research & Practice, 13*, 204-219.

Fuchs, L. S., & Fuchs, D. (2007). A model for implementing responsiveness to intervention. *Teaching Exceptional Children, 39*, 14-20.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22,* 27–48.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*, 20-28.

Gaur, A. S., & Gaur, S. S. (2006). *Statistical methods for practice and research: A guide to data analysis using SPSS*. Thousand Oaks, CA: Sage.

Glover, T. A., & DiPerna, J. C. (2007). Service delivery for response to intervention: Core components and directions for future research. *School Psychology Review*, *36*, 526-540.

Goffreda, C. T., Diperna, J., & Pedersen, J. A. (2009). Preventive screening for early readers: Predictive validity of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). *Psychology in the Schools*, *46*, 539-552.

Good, R. H., Kaminski, R. A., Cummings, K., Dufour-Martel, C., Petersen, K., Powell-Smith,... Wallin, J. (2011a). *DIBELS Next*. Retrieved from Dynamic Measurement Group website: http://www.dibels.org/

Good, R. H., Kaminski, R. A., Cummings, K., Dufour-Martel, C., Petersen, K., Powell-Smith,... Wallin, J. (2011b). *DIBELS Next assessment manual*. Retrieved from Dynamic Measurement Group website: http://www.dibels.org/

Good, R. H., Kaminski, R. A., Dewey, E. N. Wallin, J., Powell-Smith, K. A., & Latimer, R. J. (2011). *DIBELS Next technical manual*. Retrieved from Dynamic Measurement Group website: http://www.dibels.org/

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.

Greer, T., Dunlap, W. P., Hunter, S. T., & Berman, M. E. (2006). Skew and internal consistency. *Journal of Applied Psychology, 91,* 1351-1358. doi:10.1037/0021-9010.91.6.1351

Gresham, F. M. (2008). Best practices in diagnosis in a multitier problem-solving approach. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 281-294). Bethesda, MD: National Association of School Psychologists.

Hale, J. B., & Fiorello, C. A. (2004). *School neuropsychology: A practititioner's handbook*. New York, NY: Guilford Press.

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*, 753-777.

Hallahan, D. P., & Mercer, C. D. (2002). Learning disabilities: Historical perspectives. In R. Bradley, L. Danielson, & D. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 1-65). Mahwah, NJ: Lawrence Erlbaum.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under

alternative strong true score models. *Journal of Educational Measurement, 27*, 345-359.

Harris, D. N., & Herrington, C. D. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education*, *112*, 209-238.

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., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review*, *34*, 372-386.

Horn, J. L., & Catell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26*, 107-129.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York, NY: Wiley.

Hosp, M. K., & Fuchs, L. S. (2005). Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Review*, *34*, 9-26.

Howell, K. W., Hosp, J. L., & Kurns, S. (2008). Best practices in curriculum-based evaluation. In A. Thomas & J. Grimes

(Eds.), *Best practices in school psychology V* (pp. 349-362). Bethesda, MD: National Association of School Psychologists.

Hughes, C. A., & Dexter, D. D. (2011). Response to intervention: A research-based summary. *Theory Into Practice*, *50*(1), 4-11. doi:10.1080/00405841.2011.534909

Ikeda, M. J., Neessen, E., & Witt, J. C. (2008). Best practices in universal screening. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 103-114). Bethesda, MD: National Association of School Psychologists.

Inholt, C. (1991). *Read Naturally reading program*. St. Paul, MN: Read Naturally.

Individuals with Disabilities Education Improvement Act. (2004). Federal Regulations Part 300. Retrieved from http://pattan.net-website.s3.amazonaws.com/images/file/2011/08/15/sidebyside021209.pdf

Isaacson, W. (2009, April). How to raise standards in America's schools. *Time, 173*(16), 32-36.

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

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children*, *59*, 421-432.

Kaminski, R., Cummings, K. D., Powell-Smith, K. A., & Good, R. H. (2008). Best practices in using Dynamic Indicators of Basic Early Literacy Skills for formative assessment and evaluation. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 1181-1203). Bethesda, MD: National Association of School Psychologists.

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.

Kingsbury, G. G., Olson, A., Cronin, J., Hauser, C., & Houser, R. (2003). *The state of state standards: Research investigating proficiency levels in fourteen states*. Lake Oswego, OR: Northwest Evaluation Association.

Kovaleski, J. F. (2007) Response to intervention: Considerations for research and systems change. *School Psychology Review, 36*, 638-646.

Kovaleski, J. F., & Pedersen, J. A. (2008). Best practices in data analysis teaming. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 115-130).

Bethesda, MD: National Association of School Psychologists.

Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher, 31*, 3-12.

Lee, W. (2008). *Classification consistency and accuracy for complex assessments using item response theory* (Report No. 27). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment. Retrieved from http://www.education.uiowa.edu/centers/docs/casma-research/27casmareport.pdf?sfvrsn=0

Lichtenstein, R. (2008) Best practices in the identification of learning disabilities. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 295-218). Bethesda, MD: National Association of School Psychologists.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32*, 179-197.

McCardle, P., & Chhabra, V. (Eds.). (2004). *The voice of evidence in reading research*. Baltimore, MD: Paul H. Brookes.

McGlinchey, M. T., & Goodman, S. (2008). Best practices in implementing school reform. In A. Thomas & J. Grimes

(Eds.), *Best practices in school psychology V* (pp. 983-994). Bethesda, MD: National Association of School Psychologists.

McGlinchey, M. T., & Hixson, M. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review*, *33*, 193-203.

McKenna, M. C., & Stahl, S. A. (2003). *Assessment for reading instruction*. Washington, DC: American Psychological Association.

Medina, J., & Riconscente, M. M. (2006). Accounting for quality. *The Journal of Education, 186*(3), 3-10.

Melby-Lervåg, M., & Hulme, C. (2012). Is working memory training effective? A meta-analytic review. *Developmental Psychology*. Advance online publication. doi:10.1037/a0028228

Merino, K., & Beckman, T. (2010). Using reading curriculum-based measurements as predictors for the Measure Academic Progress (MAP) standardized test in Nebraska. *International Journal of Psychology: A Biopsychosocial Approach/Tarptautinis Psichologijos Zurnalas: Biopsichosocialinis Poziuris*, (6), 85-98.

Meyer, M. S. (2000). The ability–achievement discrepancy: Does it contribute to an understanding of learning disabilities? *Educational Psychology Review*, *12*, 315-337.

National Center for Education Statistics. (2012). *School district demographics system*. Retrieved from http://nces.ed.gov/globallocator/

National Institute of Child Health and Human Development. (2000). Report of the National Reading Panel (NIH Publication No. 00-4754). Washington, DC: Government Printing Office.

Neter, J., Kutner, M. H., Nachssheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models.* New York, NY: McGraw-Hill.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2001). Washington, DC: U.S. Department of Education.

Paleologos, T. M., & Brabham, E. G. (2011). The effectiveness of DIBELS oral reading fluency for predicting reading comprehension of high-and low-income students. *Reading Psychology*, *32*, 54-74. doi:10.1080/02702710903341262

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.

Pennsylvania Department of Education (2010). *2010-2011 Assessment handbook*. Retrieved from Pennsylvania Department of Education website:

http://www.portal.state.pa.us/portal/server.pt/community/pennsylvania_system_of_school_assessment

Pennsylvania Department of Education. (2011). *Special education statistical summary: 2010-2011*. Retrieved from http://penndata.hbg.psu.edu/documents/PennDataBooks/Statistical_Summary_2010-2011.pdf

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.

Powell-Smith, K. A., Good, R. H., Latimer, R. J., Dewey, E. N., & Kaminski, R. A. (2011). *DIBELS Next benchmark goals study* (Tech. Report No. 11). Eugene, OR: Dynamic Measurement Group.

Proximity (2013). *Pennsylvania school district demographic characteristics*. Retrieved from http://www.proximityone.com/sd_pa.htm

Restori, A. F., Gatz, G. S., & Lee, H. B. (2009). A critique of the IQ / achievement discrepancy model for identifying specific learning disabilities. *Europe's Journal of Psychology, 4*, 128-145.

Reyna, V. F., (2004) Why scientific research? The importance of evidence in changing educational practice. In P. McCardle & V. Chhabra (Eds.), *The voice of evidence in*

reading research (pp. 47-58). Baltimore, MD: Paul H. Brookes.

Roid, G. (2003). *Stanford-Binet Intelligence Scale: Fifth Edition*. Itasca, IL: Riverside.

Rutter, M., & Yule, W. (1975). The concept of specific reading retardation. *Journal of Child Psychology and Psychiatry, 16*, 181-197.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego, CA: Author.

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

Shapiro, E. S. (2008). Best practices in setting progress monitoring goals for academic skill improvement. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 114-158). Bethesda, MD: National Association of School Psychologists.

Shapiro, E. S., Keller, M., Lutz, J., Santoro, L., & Hintze, J. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment*, *24*, 19-35.

Shapiro, E. S., Solari, E., & Petscher, Y. (2008). Use of a measure of reading comprehension to enhance prediction on the state high stakes assessment. *Learning and Individual Differences*, *18*, 316-328.

Shaw R., & Shaw D. (2002). *DIBELS oral reading fluency-based indicators of third grade reading skills for Colorado State Assessment Program (CSAP)*. Eugene: University of Oregon.

Sheskin, D. (2004). *Handbook of parametric and nonparametric statistical procedures (*3rd ed.). Boca Raton, FL: CRC Press.

Shinn, M. R. (2008). Best practices in using curriculum-based measurement in a problem solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 243-362). Bethesda, MD: National Association of School Psychologists.

Shinn, M. R., & Garmin, G. (2006). *AIMSWeb.* Eden Prairie, MN: Edformation.

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.

Shinn, M. R., Shinn, M. M., Hamilton, C., & Clarke, B. (2002). Using curriculum-based measurement in general education

classrooms to promote reading success. In M. Shinn, H. Walker, & G. Stoner (Eds.). *Interventions for academic and behavior problems II: Preventative and remedial strategies* (pp. 113-142). Bethesda, MD: National Association of School Psychologists.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin, 138*, 628-654. doi:/10.1037/a0027473.

Silberglitt, B. (2008). Best practices in using technology for data-based decision making. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 1869-1884). Bethesda, MD: National Association of School Psychologists.

Silberglitt, B., Burns, M. K., Madyun, N. H., & Lail, K. E. (2006). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. *Psychology in the Schools*, *43*, 527-535. doi:10.1002/pits.20175

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.

Silberglitt, B. & Hintze, J.M. (2007). How much growth can we expect? A conditional analysis of R-CBM growth rates by level of performance. *Exceptional Children 74*, 71-84.

Speece, D. (2002). Classification of learning disabilities: Convergence, expansion, and caution. In R. Bradley, L. Danielson, & D. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 467-519). Mahwah, NJ: Lawrence Erlbaum.

Stage, S., & Jacobsen, M. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review*, *30*, 407-419.

Stahl, S. A. (2004). What do we know about Fluency? Findings of the National Reading Panel. In P. McCardle & V. Chhabra (Eds.), *The voice of evidence in reading research* (pp. 187-212). Baltimore, MD: Paul H. Brookes.

Stewart, L. H., & Kaminski, R. (2002) Best practices in developing local norms for academic problem solving. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 737-752). Bethesda, MD: National Association of School Psychologists.

Stewart, L. H., & Silberglitt, B. (2008). Best practices in developing academic local norms. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 225-

242). Bethesda, MD: National Association of School Psychologists.

Stuebing, K. K., Fletcher, J. M., Branum-Martin, L., & Francis, D. J. (2012). Evaluation of the technical adequacy of three methods for identifying specific learning disabilities based on cognitive discrepancies. *School Psychology Review*, *41*, 3-22.

Success for All Foundation. (2007). *4Sight Benchmark Assessments.* Retrieved from http://www.successforall.org

Tilly, W. D. (2006). Response to intervention: An overview. What is it? Why do it? Is it worth it? *The Special Edge 19*(2), 1, 4, 5, 10.

Tilly, W. D. (2008). The evolution of school psychology to science-based practice: Problem solving and the three-tiered model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 17-36). Bethesda, MD: National Association of School Psychologists.

Torgesen, J. K. (2002) Empirical and theoretical support for direct diagnosis of learning disabilities by assessment of intrinsic processing weaknesses. In R. L. Danielson & D. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 565-613). Retrieved from EBSCOhost.

Torgesen, J. K., Alexander, A., Wagner, R., Rashotte, C., Voeller, K., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities*, *34*, 33-58.

Torgesen, J. K., & Hudson, R. (2006). Reading fluency: Critical issues for struggling readers. In S. J. Samuels and A. Farstrup (Eds.), *Reading fluency: The forgotten dimension of reading success.* Retrieved from http://www.fcrr.org/publications/publicationspdffiles/Fluency_chapter-Torgesen%26Hudson.pdf

Upah, K. R. F. (2008). Best practices in designing, implementing, and evaluating qualify interventions. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 209-223). Bethesda, MD: National Association of School Psychologists.

VanDerHayden, A. M. (2011). Technical adequacy of response to intervention decisions. *Exceptional Children,* 77, 335-350.

Wayman, M. M., Wallace, T., Wiley, I. H., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education*, *41*, 85-120.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children: Fourth Edition*. San Antonio, TX: Psychological Corporation.

Wiley, H., & Deno, S. L. (2005). Oral reading and maze measures as predictors of success for English learners on a state standards assessment. *Remedial & Special Education*, *26*, 207-214.

Wilson, B. (1998). *Wilson Reading System*. Millbury, MA: Wilson Language Training.

Wise, J. C., Sevcik, R. A., Morris, R. D., Lovett, M. W., Wolf, M., Kuhn, M.,... Schwanenflugel, P. (2010). The relationship between different measures of oral reading fluency and reading comprehension in second-grade students who evidence different oral reading fluency difficulties. *Language, Speech & Hearing Services In Schools*, *41*, 340-348. doi:10.1044/0161-1461(2009/08-0093)

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III.* Itasca, IL: Riverside.

Ysseldyke, J., Burns, M. K., Scholin, S. E., & Parker, D. C. (2010). Instructionally valid assessment within response to intervention. *Teaching Exceptional Children, 42*(4), 54-61.

APPENDICES

Indiana University of Pennsylvania
www.iup.edu

Institutional Review Board for the
Protection of Human Subjects
School of Graduate Studies and Research
Stright Hall, Room 113
210 South Tenth Street
Indiana, Pennsylvania 15705-1048

P 724-357-7730
F 724-357-2715
irb-research@iup.edu
www.iup.edu/irb

September 28, 2011

Matthew R. Ferchalk
444 Hill Road
Wernersville, PA 19565

Dear Mr. Ferchalk:

Your proposed research project, "Test nationally, benchmark locally: Using local *DIBELS* benchmarks to predict performance on the PSSA," (Log No. 11-220) has been reviewed by the IRB and is approved as an expedited review for the period of September 28, 2011 to September 28, 2012.

It is also important for you to note that IUP adheres strictly to Federal Policy that requires you to notify the IRB promptly regarding:

1. any additions or changes in procedures you might wish for your study (additions or changes must be approved by the IRB before they are implemented),
2. any events that affect the safety or well-being of subjects, and
3. any modifications of your study or other responses that are necessitated by any events reported in (2).

Should you need to continue your research beyond September 28, 2012 you will need to file additional information for continuing review. Please contact the IRB office at (724) 357-7730 or come to Room 113, Stright Hall for further information.

Although your human subjects review process is complete, the School of Graduate Studies and Research requires submission and approval of a Research Topic Approval Form (RTAF) before you can begin your research. If you have not yet submitted your RTAF, the form can be found at http://www.iup.edu/page.aspx?id=91683 .

This letter indicates the IRB's approval of your protocol. IRB approval does not supersede or obviate compliance with any other University policies, including, but not limited to, policies regarding program enrollment, topic approval, and conduct of university-affiliated activities.

I wish you success as you pursue this important endeavor.

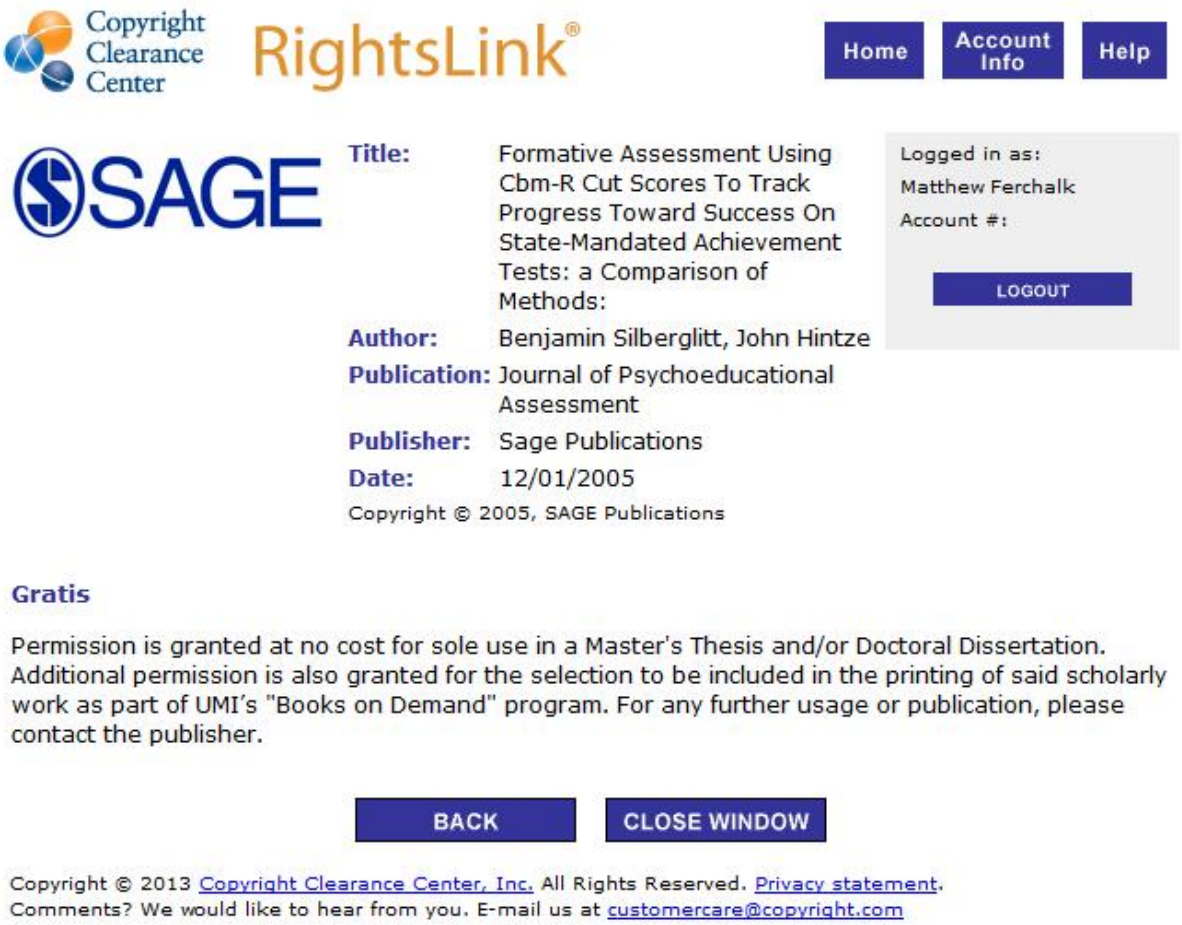
Sincerely,

John A. Mills, Ph.D., ABPP
Chairperson, Institutional Review Board for the Protection of Human Subjects
Professor of Psychology

JAM:jeb

xc: Dr. Timothy Runge, Dissertation Advisor
Ms. Jean Serio, Secretary

Appendix B, Letter of Permission from Sage Publications



Copyright Clearance Center RightsLink®

Home | Account Info | Help

SAGE

Title: Formative Assessment Using Cbm-R Cut Scores To Track Progress Toward Success On State-Mandated Achievement Tests: a Comparison of Methods:

Author: Benjamin Silberglitt, John Hintze

Publication: Journal of Psychoeducational Assessment

Publisher: Sage Publications

Date: 12/01/2005

Copyright © 2005, SAGE Publications

Logged in as:
Matthew Ferchalk
Account #:

LOGOUT

Gratis

Permission is granted at no cost for sole use in a Master's Thesis and/or Doctoral Dissertation. Additional permission is also granted for the selection to be included in the printing of said scholarly work as part of UMI's "Books on Demand" program. For any further usage or publication, please contact the publisher.

BACK | CLOSE WINDOW

Copyright © 2013 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com