The Decision-making Utility & Predictive Power of DIBELS for Students' Reading Achievement in Pennsylvania's Reading First Schools

by

Amanda M. Kloo

Bachelor of Arts, Chatham College, 1998

Master of Arts, Columbia University, 1999

Submitted to the Graduate Faculty of

The School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2006

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Amanda M. Kloo

It was defended on

November 22, 2006

and approved by

Rita M. Bean, PhD, Professor
Department of Instruction and Learning, University of Pittsburgh

Rollanda O'Connor, PhD, Professor
Graduate School of Education, University of California Riverside

George J. Zimmerman, PhD, Associate Professor
Department of Instruction and Learning, University of Pittsburgh

Audrey T. Kappel, PhD, Assistant Research Professor
Department of Instruction and Learning, University of Pittsburgh

Dissertation Advisor: Naomi Zigmond, PhD, Professor
Department of Instruction and Learning, University of Pittsburgh

The Decision-making Utility & Predictive Power of DIBELS for Students' Reading Achievement in Pennsylvania's Reading First Schools

Amanda M. Kloo, PhD

University of Pittsburgh, 2006

The purpose of this study was to examine the predictive strength and decision-making utility of the *Dynamic Indicators of Early Literacy Skills* (DIBELS). Specifically, the study examined whether DIBELS benchmarks correctly differentiated among students who were at-risk for reading failure and those who were not as measured by end-of-third-grade achievement on the Pennsylvania System of School Assessment (PSSA) in Reading First schools. More broadly, this study addressed the effectiveness of DIBELS for early identification of children considered to be at-risk for reading failure using the author-recommended benchmarks. Additionally, data were analyzed to determine whether first grade cut-points were appropriately sensitive and specific in relation to long-term predictions (end of third grade) of special education status. When comparing within-year achievement trends, results indicated that DIBELS was generally predictive of first through third grade students' Fall to Spring achievement. However, some students did demonstrate erratic achievement. Receiver Operating Characteristic (ROC) analyses revealed that the author-recommended cut-points for the Fall subtests resulted in concerning numbers of false negative and false positive predictions of reading achievement. In fact, the cut-points for the phoneme segmentation fluency (PSF) subtest were found to have a statistically inappropriate balance of sensitivity and specificity. Hierarchical Linear Modeling (HLM) analyses of students' long-term achievement showed that the DIBELS measures

administered early in first grade were generally not predictive of third grade reading achievement for students in these Reading First schools. In fact, first grade results explained only 18% of the variability in students' third grade reading scores on the PSSA. Finally, logistic regression results suggest that students' socio-economic status and race were more accurate predictors of end-of-third grade special education status than their first grade reading achievement on the DIBELS. The overall limited predictive value of DIBELS on students' long-term reading achievement raises important concerns about over-reliance on DIBELS in an early intervention framework like Pennsylvania's Reading First initiative and in school-wide educational decision making systems such as Response-to-Intervention (RTI).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

As I come to the end of this journey and embark on many new adventures, I am thankful for this opportunity to recognize those who have traveled with me and to reflect on the invaluable impact of their personal and professional support and guidance. Above all, I would like to thank my husband, Rob, for his love, support, and patience throughout this process and in all of my life's journeys. His unconditional caring and confidence were the inspiration that encouraged me to forge onward even when faced with roadblocks in the path. I would also like to thank my family for their ceaseless love, faith, and limitless support-- especially my mother who is the strength, fortitude, and determination behind every step that I take (literally). Also, thank you, Dana and the many friends who have cheered me around each new bend and supported me when I stumbled off course.

I would like to thank everyone in the Department of Instruction and Learning (faculty, students, and staff alike) for three and a half years filled with incredible support, invaluable education, and everlasting friendships. To my fellow doctoral students, Jodi, Kim, Karen, Lynn, and Louise-- thank you for walking this path by suffering through the challenges and celebrating the accomplishments with me along the way. To my dear friend, Tommy—thank you for statistical and technological brilliance, never-ending support, and enduring friendship. I would like to also thank the many students who have touched my life and helped me to become a more compassionate nurturer, a stronger teacher, and a more knowledgeable researcher. My sincerest

thanks goes to Dr. Rita Bean, Dr. George Zimmerman, and Dr. Audrey Kappel for their time, expertise, and guidance throughout this process. I would also like to thank Dr. Rollanda O'Connor for sharing her extensive knowledge, limitless enthusiasm, and valuable wisdom with me. Finally, I would like to extend my deepest appreciation to my Advisor, Dr. Naomi Zigmond, for sharing her wealth of knowledge and experience with me and affording me innumerable opportunities to learn and grow in the field. Your mentorship and friendship made this journey possible.

1.0 CHAPTER ONE

1.1 INTRODUCTION

The nation is faced with an educational epidemic—too many children suffer from chronic reading failure. Data published by the National Assessment of Educational Progress indicates that over 62% of fourth graders read below proficient levels and that over 38% of those struggling students read below basic levels of performance (NAEP, 2005). Proficient reading skills are critical not only to school success, but also to successful post school outcomes. Students with poorly developed basic reading skills in the primary grades are placed at a significant disadvantage in subsequent grades. Empirical evidence suggests that third grade, and possibly second, is too late for classroom instruction to have a significant impact on reading acquisition (Chard & Kame'enui, 2000). There is .88 probability that students identified as poor readers at the end of first grade will also be identified as poor readers at the end of fourth grade (Juel, 1988). Similarly, Juel noted that students identified as average readers at the end of first grade have a .87 probability of remaining average readers at the end of fourth grade. More recently, results from a study conducted by Felton and Wood (1992) indicated that children who failed to demonstrate strong reading skills in first grade had a 90% chance of remaining poor readers, even after 3 subsequent years of instruction. These findings imply that children's reading trajectories are established early and remain stable across grade levels and time.

However, research suggests that reading failure can be prevented if identified and treated early (National Institute of Child Health and Human Development, 2000; National Research Council, 1998). For example, research reveals that the reading performance of low-performing and at-risk students can be raised to grade-level expectations within the first 3 years of school with appropriate intervention (Chard & Kame'enui, 2000; Coyne, Kame'enui, & Simmons, 2001; Good, Simmons, & Kame'enui, 2001; Torgeson, 2000; Torgeson, et al., 2001). Moreover, this increase in performance can be sustained throughout successive grade levels; therefore, identification and intensive intervention early in a student's career can positively alter established reading trajectories (National Research Council, 1998; Torgeson, et al., 2001). Consequently, it is imperative that children at-risk are identified early and receive appropriately developed, intensive, and timely interventions.

Timely identification depends however, on valid and reliable assessment measures of core reading skills that are predictive of reading achievement and can guide the development of high intensity interventions in the classroom. Recent legislation such as No Child Left Behind (NCLB, 2001), the Reading First Initiative (2001), and the Individuals with Disabilities Education Improvement Act (2004), charge schools with the daunting task of developing school-wide reading assessment and intervention systems beginning in early grades to prevent reading failure from taking hold. These formative evaluation procedures should help to screen students for reading failure, identify specific reading skill deficits, monitor student progress, inform instructional practice, and assist in making eligibility decisions for special education services.

The purpose of this study is to examine the predictive strength and decision-making utility of the Dynamic Indicators of Early Literacy Skills (DIBELS; Good & Kaminski, 2002), a general outcome measure for reading achievement used in Pennsylvania Reading First schools.

Specifically, the study will examine whether DIBELS benchmarks correctly differentiate among students who are at-risk for reading failure and those who are not as measured by end-of-grade achievement on the Pennsylvania System of School Assessment (PSSA). More broadly, this study will address the effectiveness of DIBELS for informing educational decisions with regard to current benchmarks for identification of children considered to be at-risk. Additionally, 3rd grade special education referral data will be analyzed to determine whether DIBELS scores are appropriately sensitive and specific in relation to special education eligibility determination.

1.2 REVIEW OF LITERATURE

1.2.1 Foundational Reading Skills

Research suggests that the differentiating factor for successful versus unsuccessful readers is foundational skill knowledge (Kaminski & Good, 1996; Torgeson, Wagner, & Rashotte, 1997; Yopp, 1988). Multiple sources of reading research agree that not only are the skills involved in early reading acquisition critically important to the ability to comprehend text and future reading success, but also are the skills that prove to be most troublesome for students with reading disabilities and significant reading difficulties (Adams, 1990; CIERA, 2001; NRC, 1998, Stanovich & Stanovich, 1995). The National Reading Panel (NRP, 2000) reports that those stepping stone literacy skills include: phonological awareness and phonemic awareness, alphabet principle (phonics), and fluency. The Panel recommended that each of these skills be addressed daily during reading instruction to ensure that young readers embark on a path to proficient reading.

1.3 PHONOLOGICAL AND PHONEMIC AWARENESS

While often used synonymously, the terms "phonological" and "phonemic awareness" refer to similar but distinctive skills. Phonological awareness is the more encompassing term defined by Kame'enui and Carnine (1998) as a range of activities in which individuals manipulate either individual or groups of sounds. Phonemic awareness is a sub-component of phonological awareness that focuses specifically on recognizing and manipulating individual sounds. Phonemic awareness is defined as "an awareness of the phonological segments in speech—the segments that are more or less represented by alphabetic orthography" (Blachman, 2000, p. 483). The NRP categorizes phonemic awareness as the ability to manipulate the smallest units constituting spoken language. Instruction in the awareness of sounds as well as the ability to manipulate those sounds appears to be a crucial stepping stone for later instruction in and mastery of alphabetic awareness, especially for children with reading deficits (Adams, 1990; Blachman, 1997; O'Connor & Jenkins, 1999; Shaywitz, 1996; Stanovich, 1992, Wagner & Torgeson, 1987). Therefore phonemic and phonological awareness skills are hierarchically related to more advanced reading skills.

Research also suggests that the relationship between phonemic awareness and literacy achievement is reciprocal (Lundberg, 1991; Perfetti, Beck, Bell, Hughes, 1987). That is, on the one hand, phonemic awareness aids in later word recognition which promotes advanced reading skills. On the other hand, advanced reading skills promote more reading, which in-turn promotes stronger phonemic awareness, which ultimately promotes even greater gains in reading. Special education research suggests that this cycle is interrupted for students with dyslexia and severe reading difficulties. Deficits in phonemic and phonological awareness are common factors for

these children that result in acute, all encompassing reading impairments. Unless detected early and addressed in explicit instruction, this missing link will negatively impact reading progress not only in elementary grades but also in adolescence and adulthood (Wolf & Bowers, 1999; Fletcher et al., 1997). Considering this link, it is evident that children who begin their reading instruction with greater levels of phonemic awareness have a "powerful bootstrapping mechanism to reading progress" (Stanovich, 1992, p. 308). Overall, it seems that children who lack adequate phonological skills are likely to be the poorest readers in school. Multiple studies have shown that phonemic awareness is one of the best predictors of how well children will read during the primary grades (Juel, 1988; Lyon, 1999; NRP 2000; Share, Jorm, Maclean & Matthews, 1984).

1.3.1 Alphabetic Understanding

Alphabetic understanding involves the ability to connect letter recognition with phonological awareness, to match the textual presentation of a word to the aural presentation of a word. For example, when encountering the word *cat*, alphabetic understanding comes in to play when the reader decodes the word by recognizing that the letter *c* makes the /k/ sound, *a* makes the /ă/ sound, etc. This process results in fully recoding the word "*cat*". The NRP (2000) and the Center for the Improvement of Early Reading Achievement (CIERA, 2001) refer to alphabetic understanding as "phonics." A reader's ability to apply phonics skills by mapping print to speech is crucial to reading success.

Ehri (1992) provides extensive evidence that successful reading comprehension is dependent on strong phonics ability. When readers learn to decode words, they initially isolate and segment the subcomponents of printed words by sounding out letter by letter. As their

alphabetic understanding and application of those phonics skills strengthen competent decoders jump to recognizing common letter chunks such as prefixes, suffixes, blends, and rimes which lead to the ability to accurately blend those chunks into words or recode them. Efficient decoding and recoding skills contribute to automatic word recognition skills which influence the reader's ability to accurately and quickly read words in text leading to improved understanding. In a landmark longitudinal study, Bradley and Bryant (1983) studied the effects of directly teaching 4 and 5 year-old students letter-sound connections. They found that those students demonstrated strong reading skills three and four years later.

Numerous studies provide evidence that identifying alphabetic code deficits and providing direct instruction in phonics are essential components to skilled reading (Beck & Juel, 1995; Chall, 1989; Ehri, 1991; NRP, 2000; Share & Stanovich, 1995). Even more impressive are results that indicate that interventions that address both phonological awareness and phonics skills have sustained positive effects for several years. Specifically, at-risk kindergarten students receiving explicit code-based instruction improved reading outcomes by second grade (Vellutino et al., 1996). Moreover, Torgeson and others (1997) found reading acquisition was more successful for poor readers and students with severe reading disabilities when comprehensive instruction in phonological awareness and phonics was provided.

1.3.2 Fluency

The Center for the Improvement of Early Reading Achievement (CIERA, 2001) reports that fluency is paramount because it provides a bridge between phonemic awareness, alphabetic understanding, and comprehension—the end goal of reading. In 1980, Schreiber described reading fluency as "that level of reading competence at which textual material can be

effortlessly, smoothly and automatically understood" (p. 177). Similarly, Meyer and Felton (1999) categorized it as "the ability to read connected text rapidly and effortlessly with little conscious attention to the mechanics of reading such as decoding" (p. 284). Skilled readers activate decoding skills, semantic knowledge, and background knowledge efficiently to make connections between words, sentences, and paragraphs. They then relate these connections to broader ideas to understand story plots or informational topics. The coordination of all of these thought processes translates to smooth and accurate oral reading. This reading ease, supported by the aforementioned sub-skills of reading, is the core of the 2000 National Reading Panel's (NRP) definition of fluency as "the immediate result of word recognition proficiency" (chap. 3, p. 9).

Research indicates that non-fluent ("dysfluent") oral reading strongly correlates with poor reading comprehension (Fuchs, Fuchs, Hosp, & Jenkins, 2001). Because dysfluent readers struggle to identify and produce words as well as process text at laborious rates, they have difficulty comprehending text. Reading skill is acquired through the mastery of subcomponents of reading beginning with immediate recognition of letters, moving toward immediate recognition of words, resulting in immediate recognition of phrases and sentences, etc. LaBerge and Samuels (1974) state that these lower-level cognitive processes must be in place before higher-level cognitive processing is possible. Only when those foundational subskills are mastered and decoding occurs automatically, can comprehension occur. Struggling readers are forced to expend large amounts of cognitive capacity on word identification and therefore have little cognitive capacity left for comprehension (Meyer & Felton, 1999; Perfetti, 1985,). They are unable to progressively and selectively store information into different cognitive "slots" for later retrieval because all slots are filled word by word with no movement from short-tem,

memory, to working memory, to long-term memory. For example, when reading a narrative paragraph, fluent readers continually free-up cognitive space as they process text and store information about characters in one slot, plot in another, resolution in yet another and so on, to ultimately understand the story grammar of the passage. In contrast, dysfluent readers sacrifice an extraordinary amount of cognitive capacity focusing on recognizing and verbalizing individual text units. One slot is filled with the first ten words of the story, another with the next ten, and yet another with the next ten etc. This process quickly overloads working memory and prevents understanding of content (Perfetti, 1985). The result of this heavy cognitive load is choppy, slow, iterations of disconnected words and thoughts that resemble a vocabulary list more than a story. As reading fluency improves the reader's focus shifts from a word-reading level to a sentence-reading level and cognitive capacity is allocated more appropriately, usually resulting in stronger comprehension.

1.4 ASSESSMENT OF FOUNDATIONAL READING SKILLS

While the aforementioned skills do not capture all components of skilled reading, researchers agree that they do represent the critical skills that should be targeted in early elementary school. Systematic assessment, instruction, and intervention on these skills help to ensure successful reading outcomes. Given the climate of increased accountability in education and the research-proven danger of delaying intervention for struggling students, schools recognize the importance of utilizing early literacy assessment tools to improve decision-making regarding curriculum design and instructional practice.

1.4.1 Overview of Educational Decision-Making

1.4.1.1 The No Child Left Behind Act of 2001 (NCLB)

The sweeping reform legislation The No Child Left Behind Act of 2001 (NCLB) attempts to "close the achievement gap with accountability, flexibility, and choice so that no child is left behind" (P.L. 110-110, 2002). The law is based on four tenets: accountability for results, emphasis on research-based practice, expanded parental options for educational choice, and expanded local control and flexibility of education. Under this law, states are required to measure every public school students' progress in reading and math in grades three through eight annually and at least once in grades ten through twelve using assessments aligned to state academic standards. In Pennsylvania, the tool for monitoring students' progress is the Pennsylvania System of School Assessment (PSSA). The result of this high-stakes assessment speaks to each school's accountability by measuring their adequate yearly progress (AYP) toward getting their students to meet academic standards. Schools that fail to meet AYP face strict penalties such as personnel replacement and district take-over.

1.4.1.2 The Reading First Initiative

Two important features of NCLB are a focus on prevention and research-based instruction and assessment. The Reading First Initiative, Part B of NCLB legislation, focuses on providing state and local education agencies with the resources to practice highly effective reading instruction based on scientifically-based research for children in kindergarten through third grade. The overall goal is for all third graders to read at or above grade level by 2013. Assessment and accountability are hallmarks of Reading First. Schools receiving Reading First funds are required to practice systematic screening, diagnostic, and classroom-based reading

assessments. This prevention approach focuses on early intervention to alter struggling students' reading trajectories before they fall too far behind.

1.4.1.3 Progress Monitoring (PM) and Response to Intervention (RTI)

Consistent with the legal requirements discussed above, progress monitoring (PM) provides careful links between assessment and the instructional process. PM is a research-based practice used to assess students' academic performance and evaluate the effectiveness of instruction (Fuchs, Fuchs, & Hamlett, 1989). Systematic progress monitoring involves screening all students for potential reading failure, diagnosing specific skill deficits and making data-driven instructional decisions (Fuchs, Fuchs, Hamlett, & Allinder, 1991; Speece & Case, 2001).

PM has also been shown to aid in making eligibility decisions as a part of the Response-to-Intervention (RTI) framework in which student eligibility for special education services is a function of the students' non-responsiveness to effective interventions (Vaughn, Linan-Thompson, & Hickman, 2003). In most models of RTI, students are first exposed to high quality interventions and are only considered eligible for special education once they have not responded to these or more intensively focused intervention strategies. Therefore, PM has become a valuable evidentiary tool used to determine whether students are responding to high quality interventions (Speece & Case, 2001; Speece, Case, & Molloy, 2003).

1.4.1.4 High-stakes and Low-stakes Assessment

The purpose and use of assessment data are distinctly different between high-stakes and low-stakes testing (Salvia & Ysseldyke, 1998). As mentioned earlier, the PSSA is a high-stakes test. This type of evaluation involves summative measurement of a student's knowledge of discrete skills at a single point in time. In light of educational accountability and making AYP,

the results of this summative assessment lead to high-cost decision-making such as staffing, funding, promotion of grade level etc. In contrast, low-stakes assessments most often involve internal educational decisions such as instruction, grouping, curricular design etc. Formative assessments such as progress monitoring involve continuous measurement of students' mastery-level of skills and learning over time. Research has found that formative low-stakes assessments provide more reliable, valid, and explicit information about students' progress toward meeting educational goals and facilitate greater student achievement than their high-stakes counterparts (Fuchs & Fuchs, 1986).

1.5 ASSESSMENT OF FOUNDATIONAL READING SKILLS FOR EDUCATIONAL DECISION-MAKING

1.5.1 DIBELS

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS, Good & Kaminski, 2002) are a series of subtests measuring the foundational reading skills highlighted previously: phonological and phonemic awareness, alphabetic principle, and oral reading fluency. According to the authors, DIBELS can be used to a) determine children at-risk for reading failure; b) determine which children need additional instruction in specific reading skills; and c) determine whether current instruction for these children is effective. Research suggests DIBELS is a valid assessment for accurate identification of students' reading difficulties and instructional needs and is a particularly valuable tool in a problem-solving model in which students' deficits can be

remedied before they fall significantly behind their peers (Good, Kaminski, Simmons, & Kame'enui, 2001).

In light of the school-wide assessment model defined by Reading First, use of DIBELS as a progress monitoring tool has become increasingly popular. In fact, it is the formative assessment of choice in Pennsylvania Reading First schools. Because the predictive strength of this measure is explored the proposed study, it is useful to discuss research examining DIBELS' utility for impacting educational decisions.

In 2001, researchers at the University of Oregon (Good et al., 2001) reported the results of a linked longitudinal study exploring the predictive validity of the DIBELS early literacy skills subtests (e.g., Phoneme Segmentation [PSF] and Nonsense Word Fluency [NWF]) and oral reading fluency subtest (DORF) on the students' later DIBELS achievement and third grade reading competence on the Oregon Statewide Achievement Test (OSA). Within year analyses compared 353 kindergarten students' winter to spring DIBELS achievement; 378 first graders' winter to spring DIBELS achievement; and 364 third graders' spring DIBELS to spring OSA achievement during the 2000 school year. Results indicated that 91% of the kindergarteners who reached PSF goals in January also met PSF goals in the spring (r=.34). Moreover, 90% of the first grade students who reached NWF benchmark goals in the winter met DORF benchmark goals in the spring (r=.78). Finally, 96% of the third grade students who achieved proficient fluency rates on the Spring DORF also exceeded expectations on the reading subtests of the OSA.

Cross-year analyses compared 302 kindergarteners' spring DIBELS (1999) performance to their later winter DIBELS (2000) performance in first grade and 342 first graders' spring DIBELS (1999) performance to their later spring DIBELS performance in second grade (2000).

Results revealed that (55%) of the kindergarten participants who reached the DIBELS PSF benchmark in the Fall of 1999 also reached NWF benchmark goals in 1st grade in the winter of 2000 (r=.38). Moreover, 97% of the first graders who met end-of-year DORF benchmarks in 1999 later met second grade Spring DORF benchmarks one year later (r=.82). The authors concluded those successful within-year and cross-year outcomes indicated that if a student reaches benchmark on each DIBELS subtest, "the odd are in his/her favor" for successfully meeting subsequent benchmarks and future reading goals.

Those results yield positive evidence to support the predictive power of DIBELS and its use as an early screening tool. However, it should be mentioned that in the technical manual the authors note, "DIBELS are not designed to serve as a comprehensive or diagnostic reading assessment tool. Rather, they are intended to provide a fast and efficient indication of the academic well-being of students with respect to important literacy skills" (Good, Simmons, Kame'enui, Kaminski, & Wallin, 2002, p. 53).

Fuchs & Fuchs (2000) compared DIBELS' strengths and weaknesses as a progress monitoring tool to Deno's (1992) curriculum-based measurement (CBM) and Bloom's (1976) Mastery Measurement (MM). Each measure was critiqued against six "desirable" criteria including: traditional psychometric standards of reliability and validity, capacity to predict and model growth, sensitivity to individual change, independence from specific instruction, capacity to inform teaching, and feasibility. Results showed that CBM met all criteria with the exception of feasibility (unless using computer-based programs); MM met only treatment sensitivity and capacity to inform teaching; while DIBELS met only traditional psychometric standards due to an apparent lack of empirical evidence on its relationship to the other standards. The authors

cautioned against using DIBELS as a "catch all" progress monitoring assessment and called for more research on its usefulness as measure of students' growth over time.

It is important to note that research on the concurrent validity of DIBELS with other standardized measures of fluency is also lacking. Hintze, Ryan, & Stoner (2003) compared 86 kindergarteners' scores on DIBELS to their scores on the *Comprehensive Test of Phonological Processing* (CTOPP). Specific subtests measuring similar constructs of phonological awareness were administered in the winter of the kindergarten year. Receiver Operating Characteristic analysis results showed moderate to strong correlations between the two measures. However, while DIBELS was highly sensitive in identifying children with low phonological awareness skills on the CTOPP (true positives), the measure also identified many children as low performers who did not perform poorly on the CTOPP (false positives). This low specificity of the DIBELS cut- scores raises concern when considering both the widespread use of DIBELS and the impact of misappropriation of instructional support resources for students incorrectly identified by DIBELS as "at-risk" for reading failure. Clearly, further research on the diagnostic accuracy of DIBELS is needed and this need serves as impetus for the proposed study.

1.5.2 Diagnostic Accuracy

Diagnostic accuracy is a valuable feature of an assessment tool used for educational decision-making. Swets, Dawes and Monahan (2003) refer to it as an instrument's accuracy and efficiency of correctly distinguishing between two alternatives—the presence of a problem and the absence of a problem. Two key components of diagnostic accuracy are sensitivity and specificity, Sensitivity is related to a measure's true-positive prediction rate. Specificity is related to a measure's true-negative prediction rate. There is an inverse relationship between the

two so the more balanced the sensitivity and specificity, the more accurate the assessment tool (see "Definition of Terms" for more detailed descriptions of each).

A useful measure for determining the diagnostic accuracy of an instrument is Receiver Operating Characteristic (ROC) analysis, which explores the range of cut-scores for a particular measure to find the optimal ratio of true positive and false positive decisions for an adequate balance of sensitivity and specificity for specific cut-scores (Tatano-Beck & Gable, 2001). The graphic display of the range of scores is represented by the ROC Curve. The area under the ROC curve represents the diagnostic accuracy of the instrument, therefore; the optimal cut-score is generally near the shoulder of the curve. Swets (1988) recommends that .75 serve as the minimum criterion for diagnostic accuracy indicating a "fair" balance of the characteristics described above.

1.5.3 Summary

The NRP (2000) reported that "without early identification (before entry into the third grade) reading difficulties severe enough to hinder learning and the enjoyment of reading will persist into adulthood unless intensive and specialized remediation programs are provided." Effective assessment is a key aspect of early identification. Educational laws like NCLB 2001 place increasing emphasis on high-stakes achievement outcomes, program evaluation, and instructional accountability making highly predictive and accurate assessment measures invaluable tools for school improvement and increased achievement. The DIBELS has been celebrated as such a tool. However, while the field is currently deluged with research on the theoretical soundness of the reading skills measured by DIBELS, data about its predictive validity and utility as a decision making tool are noticeably lacking. As the popularity of this test

grows and the types of decisions made based on students' achievement on this test become more serious, it is important to determine whether its usefulness generalizes beyond the original Oregon study.

2.0 CHAPTER TWO

2.1 RELEVANCE OF THE STUDY

With the current political and educational emphasis on accountability, standards-based reform, and high-stakes assessment, the field of education has seen a dramatic shift toward focusing monies, resources, and personnel on improving students' reading outcomes (Ravitch, 1999). Schools are now faced with the challenge of ensuring that all children become proficient readers by the end of third grade (NCLB, 2001). Although there is much debate about how to best define and measure what proficient reading really "is", current legislation mandates that schools measure proficiency using a high-stakes/standardized standards-based assessments (Braden, 2002). However, most standardized state-wide assessments are not administered until third-grade by which time many students have fallen hopelessly behind with intractable reading trajectories (Good et al.; 2001 Torgeson, 1998). Therefore, it is imperative that reading researchers extend considerable energy into identifying valid and reliable measures for targeting students for early reading intervention.

The Dynamic Indicators of Early Literacy Skills (DIBELS) was developed as low-stakes measures of early literacy skills that could be used to predict of performance on high-stakes outcome measures (Good & Kaminski, 1996; Good et al., 2001.) Educational reform legislation has prompted increasingly widespread use of DIBELS as a diagnostic tool for identifying

students at-risk for reading difficulties. In Pennsylvania Reading First schools, DIBELS results build the foundation for instructional decision making on the basis of which teachers make grouping and resource allocation decisions including: increased instructional time, additional instructional personnel, and differentiated instructional plans. Given the weight placed on DIBELS assessment results for educational decision making and service provision, further investigation of its utility and predictive validity is needed.

2.2 SIGNIFICANCE OF THE PROBLEM

The effectiveness of school-wide educational decision making systems such as progress monitoring and Response-to-intervention (RTI) directly relates to the accuracy of the measure used to identify students at-risk for reading problems and the type of decisions being made such as group placement, intervention intensity, and referral for special education. At first glance, DIBELS high degree of sensitivity appears desirable to identify children at-risk for reading problems. That is to say that by casting a wide net, the DIBELS ensures that high-quality and intense interventions can be provided to the students who need them most. Some researchers believe that the benefit of correctly identifying this population of the neediest students outweighs the cost of misidentifying their more able peers (Swets, 1988). However, as charges for more accountability, effective prevention, and accurate early identification prompt schools to rely more heavily on DIBELS results for labeling children as "at risk" and allocating resources greater caution and consideration is needed when finding an appropriate balance between sensitivity and specificity (Hintze et al., 2003). It is imperative that researchers and educators more closely scrutinize this assessment's strengths and limitations.

2.3 RESEARCH QUESTIONS

The following research questions that will be addressed in this study and discussed in the following chapters make use of Pennsylvania Reading First data from the 2003-2004, 2004-2005, and 2005-2006 school years.

1) Are the established, author-recommended DIBELS cut-points for each benchmark period in grades one to three accurate predictors of achieving subsequent benchmark goals?
    a) Did first graders who achieved NWF and PSF benchmark goals in the Fall (2005) achieve NWF, PSF, and DORF benchmark goals in the Winter and Spring (2006)?
    b) Did second graders who achieved DORF benchmark goals in the Fall (2005) achieve DORF benchmark goals in the Winter and subsequent DORF benchmark goals in the Spring (2006)?
    c) Did third graders who achieved DORF benchmark goals in the Fall (2005) achieve DORF benchmark goals in the Winter and subsequent DORF benchmark goals in the Spring (2006)?
    d) Did third graders who achieved DORF benchmark goals in the Spring (2006) achieve "Proficient or Advanced" status on the PSSA?
    e) Will the use of ROC analysis result in the establishment of cut-points with greater predictive power including an appropriate balance of sensitivity and specificity?
2) How accurately do the DIBELS measures administered in the Fall and Winter of first grade predict third grade Spring achievement on the DORF and on the PSSA?

a) What is the relationship between 3rd grade students' benchmark achievement on the 2006 Spring DORF DIBELS and their earlier first grade (2003/2004) DIBELS achievement on the Fall PSF, Fall NWF and Winter DORF subtests?

b) What is the relationship between 3rd grade students' achievement on the 2006 PSSA Reading subtest and their earlier first grade (2003/2004) DIBELS achievement on the Fall PSF, Fall NWF and Winter DORF subtests?

c) How much of the variance can be explained by the Level 2 variables built into the HLM design?

   i) DIBELS 1st Grade Fall PSF, Fall NWF, and Winter DORF scores

   ii) School

   iii) Student minority status

   iv) Student SES

3) Are the DIBELS subtests administered in the Fall and Winter of first grade accurate predictors of eligibility for special education services in reading at the end of third grade?

   a) What is the relationship between students' eligibility status at the end of third grade (2006) and earlier first grade (2003/2004) DIBELS benchmark status of "at risk" on the Fall PSF & NWF subtests and Winter DORF subtest?

   b) What is the relationship between students' eligibility status at the end of third grade (2006) and their minority and SES status?

2.4 DEFINITION OF TERMS

DIBELS- The *Dynamic Indicators of Basic Early Literacy Skills is* a set of standardized, short-duration fluency measures designed to serve as general outcome indicators measuring students' proficiency of basic literacy skills documented to be highly related to reading proficiency. The authors maintain that DIBELS is a "valid, reliable, efficient measure that when given early in a child's beginning literacy experience serve as powerful predictors of later reading success" (Good & Kaminski, 2002).

Cut-point/ cut-score- A specified point on a scale of scores that serves as a decision- threshold. Scores at or above that point are interpreted differently from scores below that point (e.g., above= passing, below=failing). Multiple cut-scores on the DIBELS are defined for each subtest to establish performance standards for each grade-level.

DIBELS subtests-

- PSF- Phoneme Segmentation Fluency
- NWF- Nonsense Word Fluency
- DORF- DIBELS Oral Reading Fluency

Benchmark goal- a desired performance standard level established by the authors of DIBELS for Fall, Winter, and Spring assessment administrations.

DIBELS Descriptive Levels of Performance

- At Low Risk- defined by the DIBELS technical manual as the performance level at which an individual's score indicates the "odds are in favor of achieving subsequent outcomes" on later subtests.
- At Some Risk- defined by the DIBELS technical manual as the performance level at which an individual's score indicates the "odds are neither in favor of nor against achieving subsequent outcomes" on later subtests.
- At Risk- defined by the DIBELS technical manual as the performance level at which an individual's score indicates the "odds are against achieving subsequent outcomes" on later subtests.

DIBELS Instructional Recommendations

- Benchmark- defined by the DIBELS technical manual as corresponding with the "at low risk" performance level. Students are performing "at or above grade-level" so no changes to instruction are needed. Cut points suggest that approximately 80% of students would be in need of Benchmark-level instruction and considered "at low risk" for reading failure.
- Strategic- defined by the DIBELS technical manual as corresponding with the "at some risk" performance level. No clear prediction can be made about students' exact performance (i.e., 50-50 odds of meeting future goals), so "additional intervention" would be beneficial. Cut points suggest that approximately 15% of students would be in need of Strategic-level instruction and considered at some risk for reading failure.
- Intensive- defined by the DIBELS technical manual as corresponding with the "at risk" performance level. Students are performing "below grade-level" and are in need of

"substantial intervention." Cut points suggest that approximately 5% of students would be in need of Intensive-level instruction and considered "at risk" for reading failure.

Diagnostic accuracy- the ability of an instrument to distinguish between two diagnostic alternatives (the presence or absence of a condition), and to select the one that is correct.

Sensitivity- the probability that when a diagnostic status is present on a criterion measure the predictor measure will also indicate the presence of the diagnostic status. Sensitivity is the rate of true-positive identification.

Specificity- the probability that when a diagnostic status is absent on a criterion measure the predictor measure will also indicate the absence of the diagnostic status. Specificity is the rate of true-negative identification.

PSSA- Pennsylvania System of School Assessment (PSSA) Grade 3 Reading Assessment

is a standardized, criterion-referenced assessment designed to measure specific content standards for Reading including: reading independently, reading critically, and reading, analyzing, and interpreting literature.

HLM-Growth Model- A statistical analysis technique to explore individual differences in progress or rate over time by examining predictors of growth and individual growth estimates.

ROC Analysis- Receiver Operating Characteristic analysis is a statistical method for exploring the diagnostic accuracy of a test by providing the ratio of true positive/false positive and true negative/false negative decisions along with a range of all possible cut-scores for the designated assessment measure to find the best cut-score with the optimal balance of sensitivity and specificity.

ROC Curve- the visual display of data produced by the ROC analysis.

3.0 CHAPTER THREE

3.1 METHODOLOGY

Changes in educational policy have led to a surge of interest in the use of early literacy assessments designed to "flag" students at-risk for reading failure. The Pennsylvania Reading First Initiative recommends the use of the Dynamic Indicators of Early Literacy Skills (DIBELS) as such a tool. In Pennsylvania Reading First schools, DIBELS results influence a variety of educational decisions for students including: amount of reading instruction (number of minutes), intensity of intervention, and frequency of assessment. In an effort to be more "accountable" for student outcomes, DIBELS results also influence referral practices to determine students' eligibility for special education. Previous research raises concerns about relying on DIBELS cut-scores for educational decision-making because of imbalanced levels of sensitivity and specificity (Hintze, et al., 2003). Given the weight placed on DIBELS assessment results for educational decision making and service provision, further investigation of its utility and predictive validity is needed. The current study adds to the research base on its diagnostic accuracy and appropriateness as a screening and diagnostic tool for low-performing readers like those in Pennsylvania Reading First Schools.

3.2 PARTICIPANTS AND SETTING

The analysis included only "Round One" and "Round Two" Reading First schools to allow the researcher to reference 3rd grade students' second and first grade DIBELS scores to identify achievement trends across three years of Reading First data (2004-2006). Only complete data sets were included in the analysis to facilitate estimating growth over time. Students coded as having Limited English Proficiency (LEP) were not included in the analyses. Previous research suggests that this population exhibits uneven performance on oral reading fluency measures related to automatic word calling (Valencia and Buly, 2004).

The following table presents the 2005-2006 Pennsylvania Reading First demographic data for the population of students whose scores were analyzed to answer Question 1 and its sub-questions.

Table 1: 2005-2006 PA Reading First Demographic Data for Within Grade-level Comparisons of Achievement in "Round One" and "Round Two" Schools

	N	Percent Minority Status	Percent Economically Disadvantaged	Percent IEP
Grade 1	8,595	74%	80%	10%
Grade 2	7,925	77%	81%	12%
Grade 3	8,317	75%	79%	13%

The following table presents the 2004-2006 Pennsylvania Reading First demographic data for the population of students whose scores were analyzed to answer Questions 2 and 3.

Table 2: 2004-2006 PA Reading First Demographic Data for 1st to 3rd Grade Comparisons of Achievement in "Round One" and "Round Two" Schools

	N	Percent Minority Status	Percent Economically Disadvantaged	Percent IEP
Grade 3	9,685	74%	68%	14%

3.3 DESCRIPTION OF MEASURES

3.3.1 Dynamic Indicators of Basic Early Literacy Skills (DIBELS)

The DIBELS measures used in these analyses were fluency-based measures of early literacy skills designed to assess first through third grade students' competency with three of the "Five Big Ideas" of reading including: phonemic awareness (PSF), alphabetic principal (NWF), and reading fluency (DORF) (Good & Kaminski, 2002).

The measures examined in this study were:

Table 3: DIBELS Measures Analyzed

Grade	Fall	Winter	Spring
1	NWF, PSF	PSF, NWF, DORF	PSF, NWF, DORF
2	DORF	DORF	DORF
3	DORF	DORF	DORF

3.3.1.1 DIBELS Phoneme Segmentation Fluency (PSF)

PSF is an individually administered, standardized measure of phonological awareness designed to measure a student's ability to segment the sound units (phonemes) of an orally presented word with fluency. This measure is administered to students in the winter of kindergarten through the spring of first grade. Three or four phoneme words are said aloud to the student for one minute as the student is directed to verbally isolate each of the phonemes in the word. For example, the examiner would say, "Tell me the sounds in *mop*." The student would be expected to respond, "/m/ /o/ /p/" for a total of 3/3 correct phonemes. The number of correct phonemes produced in one minute is scored.

According to the administration manual, there are 20 alternate forms available with a one month alternate-form reliability of .88. The concurrent validity of the PSF subtest is .54 with the spring Readiness Cluster score of the Woodcock-Johnson Psycho-Educational Battery. The predictive validity of spring kindergarten PSF scores are .68 with the spring first grade Total Reading Cluster score of the Woodcock-Johnson Psycho-Educational Battery; .62 with winter first grade DIBELS NWF, and .62 with spring first grade DIBELS DORF. The table below displays DIBELS "Descriptive Levels of Performance" for First Grade PSF.

Table 4: First Grade PSF Descriptive Levels of Performance

Benchmark Period	Performance	Descriptor
Fall, Winter, Spring	PSF<10	At Risk
	10≤PSF≤35	Some Risk
	PSF≥35	Low Risk

(Adapted from: Good, R.H. & Kaminski, R.A., 2002)

3.3.1.2 DIBELS Nonsense Word Fluency (NWF)

NWF is a standardized and individually administered test of alphabetic understanding. This measure is administered to students in the winter of kindergarten through the fall of second grade. It is designed to assess a student's ability to recognize letter-sound correspondence and recode or blend into make-believe words. For example, the student is presented with a list of CVC and VC stimulus words (e.g., hoj, rop, en) and asked to verbally produce as many sounds as he can in one minute by isolating each sound (i.e., /h/ /o/ /j/) or by blending the sounds together into a fully recoded word (i.e., /hoj/). The student receives for all correctly produced sounds; therefore, in the above cases the student would earn a score of 3/3.

According to the technical manual, there are 20 alternate forms available with a one month alternate-form reliability of .78. The concurrent validity of the NWF subtest is .36 and .59 with the January and February Readiness Cluster scores of the Woodcock-Johnson Psycho-Educational Battery. The predictive validity of winter first grade NWF scores is .66 with the Total Reading Cluster score of the Woodcock-Johnson Psycho-Educational Battery is; .82 with

spring first grade DIBELS DORF; and .60 with spring second grade DIBELS DORF. The table below displays DIBELS "Descriptive Levels of Performance" for First Grade NWF.

Table 5: First Grade NWF Descriptive Levels of Performance

Benchmark Period	**Performance**	**Descriptor**
Fall	NWF<13	At Risk
	13≤NWF<24	Some Risk
	NWF≥24	Low Risk
Winter	NWF<30	At Risk
	30≤NWF<50	Some Risk
	NWF≥50	Low Risk
Spring	NWF<30	At Risk
	30≤NWF<50	Some Risk
	NWF≥50	Low Risk

(Adapted from: Good, R.H. & Kaminski, R.A., 2002)

3.3.1.3 DIBELS Oral Reading Fluency (DORF)

Like the other subtests discussed, DORF is standardized and individually administered. This subtest measures the fluency and accuracy with which a student reads connected text aloud. The procedures and passages for the DORF are a downward extension of the Curriculum Based Measurement (CBM) materials and guidelines developed by Deno, (1989) and the Test of Oral Reading Fluency (TORF).

According to the manual, alternate form reliability for the DORF ranges from .89-.94. Furthermore, concurrent validity statistics range from .52-.91. The table below displays DIBELS "Descriptive Levels of Performance" for First through Third Grade DORF.

Table 6: First-Third Grade DORF Descriptive Levels of Performance

Grade Level	Benchmark Period	Performance	Descriptor
1	Winter	DORF<8	At Risk
		8≤DORF<20	Some Risk
		DORF≥20	Low Risk
	Spring	DORF<20	At Risk
		20≤DORF<40	Some Risk
		DORF≥40	Low Risk
2	Fall	DORF<26	At Risk
		26≤DORF<44	Some Risk
		DORF≥44	Low Risk
	Winter	DORF<52	At Risk
		52≤DORF<68	Some Risk
		DORF≥68	Low Risk
	Spring	DORF<70	At Risk
		70≤DORF<90	Some Risk

Table 6 Continued

		DORF≥90	Low Risk
3	Fall	DORF<53	At Risk
		53≤DORF<77	Some Risk
		DORF≥77	Low Risk
	Winter	DORF<67	At Risk
		67≤DORF<92	Some Risk
		DORF≥92	Low Risk
	Spring	DORF<80	At Risk
		80≤DORF<110	Some Risk
		DORF≥110	Low Risk

(Adapted from: Good, R.H. & Kaminski, R.A., 2002)

3.3.2 Pennsylvania System of School Assessment (PSSA) Grade 3 Reading Assessment

The Pennsylvania System of School Assessment (PSSA) Grade 3 Reading Assessment is a standardized, criterion-referenced assessment designed to measure specific content standards for Reading including: learning to read independently, reading critically, and reading, analyzing, and interpreting literature. It also assesses students' mastery of grade-level academic standards set forth by the Pennsylvania Department of Education (PDE) and the PA Chapter 4 Regulations (PDE, 1999b). This group- administered test yields raw scores, percentile ranks, scaled scores, and performance level descriptors. Performance level descriptors indicate whether or not a

student's test performance, described in scaled score ranges, is advanced, proficient, basic, or below basic. The technical report for the 2005 PSSA reported that reliability coefficients for all Reading subtests exceeded .883. Scaled score ranges by performance level for Grade 3 are included in the table below.

Table 7: Performance Level Indicators for PSSA Grade 3 Reading Scaled Score Ranges

Performance Level	**Grade 3 SS Range**
Advanced	1442 and up
Proficient	1235-1441
Basic	1098-1234
Below Basic	1097 and below

(Adapted from: PDE, 2005, p. 8)

3.4 PROCEDURES

The following DIBELS/PSSA assessment data were analyzed for first, second, and third grade students in "Round One" and "Round Two" PA Reading First schools to make the within-year comparisons proposed in Question 1 and its sub-questions.

Table 8: Assessment Data Analyzed for Research Question 1

GR	YR	DIBELS PSF			DIBELS NWF			DIBELS DORF			PSSA READING		
		F	W	S	F	W	S	F	W	S	F	W	S
1	2006	X	X	X	X	X	X		X	X			
2	2006							X	X	X			
3	2006							X	X	X			X

Next, the following longitudinal DIBELS/PSSA assessment data were analyzed for third grade students in "Round One" and "Round Two" PA Reading First schools to make the across-year comparisons proposed in Question 2 and its sub-questions.

Table 9: Assessment Data Analyzed for Research Question 2

GR	YR	DIBELS PSF			DIBELS NWF			DIBELS DORF			PSSA READING		
		F	W	S	F	W	S	F	W	S	F	W	S
3	2006									X			X
1	2004	X			X				X				

Finally, the following data were analyzed for third grade students in "Round One" and "Round Two" PA Reading First schools to investigate the long term predictions proposed in Question 3 and its sub-question.

Table 10: Assessment/IEP Data Analyzed for Research Question 3

GR	YR	DIBELS PSF			DIBELS NWF			DIBELS DORF			IEP Status		
		F	W	S	F	W	S	F	W	S	F	W	S
3	2006												X
1	2004	X			X				X				

3.5 DATA ANALYSIS

Descriptive statistics will be reported for each question including: range of scores, means, and standard deviations. All results will be analyzed at the $p<.01$ significance level. Detailed data analysis procedures for each of the proposed research questions and the linked sub-questions are discussed in the following sections.

3.5.1 Research Question 1 Analyses

Research Question 1

Are the established, author-recommended DIBELS cut-points for each benchmark period in grades one to three accurate predictors of achieving subsequent benchmark goals?

Did first graders who achieved NWF and PSF benchmark goals in the Fall achieve NWF, PSF, and DORF benchmark goals in the Winter and Spring?

a) *Did second graders who achieved DORF benchmark goals in the Fall achieve DORF benchmark goals in the Winter and subsequent DORF benchmark goals in the Spring?*
b) *Did third graders who achieved DORF benchmark goals in the Fall achieve DORF benchmark goals in the Winter and subsequent DORF benchmark goals in the Spring?*
c) *Did third graders who achieved DORF benchmark goals in the Spring achieve "Proficient or Advanced" on the PSSA?*
d) *Will the use of ROC analysis result in the establishment of cut-points with greater predictive power including an appropriate balance of sensitivity and specificity?*

3.5.1.1 Correlation

Pearson product moment correlation coefficients were calculated for each DIBELS measure administered at each grade level as separate dependent variables to determine the strength of the relationships between the DIBELS subtests.

3.5.1.2 Partial Replication of Good et al. (2001) Study

Good et al. (2001) related the accuracy of DIBELS' risk classifications to the consistency of students' achievement over time. Specifically, they reported the percent of students who achieved both early and subsequent DIBELS goals (i.e., "low risk") and students who achieved neither early nor subsequent DIBELS goals (i.e., "at risk"). Achievement patterns for students in the "some risk" categories were not reported because the likelihood of either successful or unsuccessful future achievement for those students was uncertain.

Given the current study's focus on the predictive accuracy of DIBELS' benchmark classifications at each grade level, the current study not only examined the consistent achievement patterns discussed above, but also examined inconsistent DIBELS classifications by

calculating the percent of students who moved in and out of performance categories during the 2005-2006 school year (i.e., students initially "low risk" who became "at risk" or students initially "at risk" who became "low risk"). The results of these analyses prompted further exploration of the utility of each benchmark cut-score by conducting ROC Analyses.

3.5.1.3 The Receiver Operating Characteristic (ROC) Analyses

The Receiver Operating Characteristic (ROC) analyses involved two procedures—diagnostic accuracy analysis and ROC Curve analysis. Analyses of the diagnostic accuracy of the DIBELS subtests were conducted to determine the accuracy of DIBELS' classifications of students as "low risk" and "at risk". Decision matrices (Swets et al., 2003) were constructed to display the rates of true positive, false positive, true negative, and false negative identification. The following 2x2 matrix illustrates these relationships.

Outcome measure

(Spring DIBELS subtests at each grade level)

Predictor Measure (Fall DIBELS subtests at each grade level)		+	-
	+	True positive	False Positive
	-	False Negative	True Negative

Figure 1: Sample Diagnostic Accuracy Decision Matrix

True positive and true negative decisions suggest agreement and a strong predictive relationship between the predictor measure and outcome measure thereby accurately indicating either the presence or the absence of a problem. False positives and false negatives suggest disagreement and a weak predictive relationship between the predictor measure and outcome measure thereby inaccurately indicating either the presence or the absence of a problem. Initial analysis was conducted using the author recommended cut-points for each measure to determine current levels of sensitivity and specificity. The percentage of true positive vs. false positive/true negative vs. false negative identification statistics were reported for each DIBELS cut-point.

The rate of true positive and false positive decisions were of particular interest to this study in relation to the decision making utility of DIBELS to accurately flag students at risk for reading failure. If results suggested that the author-recommended cut-scores were not accurate predictors of future DIBELS performance, more appropriate cut-points were established based on the ROC score continuum providing all performance coordinates for the curve.

ROC Curve analyses were conducted to establish alternate cut-points with appropriate levels of sensitivity, specificity, and predictive power. This graphic representation showed the range of all possible cut-scores of a predictor measure to: 1) compare the diagnostic accuracy of multiple measures, 2) view the inverse relationship between sensitivity and specificity, and 3) select an optimal decision benchmark for a specific distribution of scores (Swets et al., 2003; Tatano-Beck & Gable, 2001). Ultimately, 2 ROC curves for each cut-score were evaluated--one with respect to successful outcome (i.e., the likelihood of achieving subsequent benchmark goals) and the other with respect to unsuccessful outcome (i.e., the likelihood of not achieving subsequent benchmark goals).

The test statistic for the area "under" the curve was examined to determine the predictive strength of the measure in question. Levels of greater than .75 were used as the "ideal" benchmark for true-positive and true negative identification resulting in decision making criteria that are both sensitive and specific in accurately identifying students at-risk for reading difficulties (Swets, 1998).

3.5.2 Research Question 2 Analyses

Research Question 2

How accurately do the DIBELS measures administered in the Fall and Winter of first grade predict third grade Spring achievement on the DORF?

a) *What is the relationship between 3rd grade students' benchmark achievement on the 2006 Spring DORF DIBELS and their earlier first grade (2003/2004) DIBELS achievement on the Fall PSF, Fall NWF and Winter DORF subtests?*

b) *What is the relationship between 3rd grade students' achievement on the 2006 PSSA Reading subtest and their earlier first grade (2003/2004) DIBELS achievement on the Fall PSF, Fall NWF and Winter DORF subtests?*

c) *How much of the variance can be explained by the Level 2 variables built into the HLM design?*

   i) *DIBELS 1st Grade Fall PSF, Fall NWF, and Winter DORF scores*

   ii) *School*

   iii) *Student minority status*

   iv) Student SES

3.5.2.1 First Grade to Third Grade DIBELS Achievement Patterns

Although Good et al. (2001) did not examine cross-year relationships between students' first to third grade performance on DIBELS or first grade DIBELS to third grade high-stakes test performance, these comparisons were of particular interest to this study. The proportion of students achieving "low risk" and "at risk" status on the specified subtests administered at the beginning of first grade and at the end of third grade were calculated. The percent of students changing risk categories were also calculated. The same comparisons were made for first grade DIBELS to third grade PSSA achievement.

3.5.2.2 Correlation

Pearson product moment correlation coefficients were calculated to determine the strength of the relationships between designated DIBELS subtests and the Reading subtest of the 3rd Grade PSSA. The resulting correlation coefficients also informed the entry order of the HLM models.

3.5.2.3 HLM Growth Modeling

The present study used the random coefficient growth modeling technique of Hierarchical Linear Modeling (HLM) to analyze the data collected within groups for Question 2 and its sub-questions. HLM helped to provide an accurate measure of change and growth because the multiple-time-point analysis accounted for individual variability and under/over estimation of observed relationships (Raudenbush & Bryk, 2002). The model was particularly informative when examining predictive validity and helped to ensure accurate prediction from DIBELS literacy measures to high-stakes assessment results for students in Reading First

schools. Furthermore, a clear advantage of applying HLM analysis to this longitudinal data was the ability to handle missing data (Bryk & Raudenbush, 1992).

The data for this study was represented by Bryk & Raudenbush's (1992) two-level growth analysis model. Level one variables represented outcome variables including students' 3rd grade performance on the Spring DORF and reading subtest of the PSSA (2006). Analysis of these variables produced individual growth estimates representing achievement outcomes. Level two variables included the predictor variables of first grade Fall PSF, NWF and Winter DORF results (2004), which represented the characteristics that explained the differences among growth trajectories. Level two variables also included the between subject contextual factors that may have acted as predictors such as: school, student minority status, and student socio-economic status to further explain variability in achievement outcomes.

Level 2 variables were arranged in the following models for analyses of impact on Level 1 variables:

Table 11: Growth Models- Outcome Achievement v. Predictor Variables

Model #	Outcome Variable	Predictor Variable
1	Spring 3rd DIBELS DORF (2006)	• Fall 1st Grade DIBELS PSF (2004) • Fall 1st Grade DIBELS NWF (2004) • Winter 1st Grade DIBELS DORF (2004) • School • Student Minority Status • Student SES
2	Spring 3rd PSSA Reading (2006)	• Fall 1st Grade DIBELS PSF (2004) • Fall 1st Grade DIBELS NWF (2004) • Winter 1st Grade DIBELS DORF (2004) • School • Student Minority Status • Student SES

3.5.3 Research Question 3 Analyses

Research Question 3

Are the DIBELS subtests administered in the Fall and Winter of first grade accurate predictors of eligibility for special education services in reading at the end of third grade?

a) *What is the relationship between students' eligibility status at the end of third grade and earlier first grade DIBELS benchmark status of "at risk" on the Fall PSF & NWF subtests and Winter DORF subtest?*

b) *What is the relationship between students' eligibility status at the end of third grade and their minority and SES status?*

3.5.3.1 Correlation

Pearson product moment correlation coefficients were calculated to determine the strength of the relationships among first grade "at risk" status on Fall/Winter DIBELS subtests, SES, minority status, and end-of-third grade special education status.

3.5.3.2 Consistency of DIBELS Classification and Special Education Status

To explore the relationship between students' first grade DIBELS risk status and eventual eligibility for special education, the following comparisons were made: 1) the percent of students who achieved DIBELS benchmarks in the first grade and were not eligible for special education in third grade; 2) who did not achieve DIBELS benchmarks in first grade and were eligible for special education in third grade; 3) who achieved first grade DIBELS benchmarks but were eligible for special education in third grade; or 4) who did not achieve first grade benchmarks but were not eligible for special education in third grade.

3.5.3.3 The Receiver Operating Characteristic (ROC) Analyses

Only the diagnostic accuracy analysis of DIBELS accuracy of third grade special education identification was conducted. The results report the number of true/false positive and true/false negative classifications based on first grade DIBELS "at risk" status. A ROC Curve analysis was not conducted for Question 3 as the sensitivity/specificity balance of the "at risk" cut points were examined in Question 1.

3.5.3.4 Logistic Regression

Logistic regression predicts a discrete outcome such as group membership. For this data, the discrete outcome of interest was end-of-third grade special education status. Tabachnick and

Fidell (1996) suggest that logistic regression is best suited for cases when the dependant variable is dichotomous such as Yes/No, 1/0 etc., while the independent variables are nominal, ordinal, ratio or interval. Grimm and Yarnold (1995) also recommend using logistic regression over other multivariate methods such as discriminant function analysis when working with large samples because it is more flexible in its assumptions and does not require the data to be normally distributed or have equal variances within groups.

Given the large sample size and dichotomous nature of the dependent variable (3^{rd} grade IEP= 1, 3^{rd} Grade No IEP= 0) logistic regression was used to calculate probability of students being identified for special education in third grade based on the first grade variables of DIBELS "at risk" status on the Fall PSF, Fall NWF, and Winter DORF subtests, minority status, and socio economic status.

4.0 CHAPTER FOUR

4.1 RESULTS

The purpose of the study was to investigate the decision-making utility and predictive strength of the DIBELS for student outcomes in Pennsylvania Reading First schools including: DIBELS benchmark achievement, end-of third grade proficiency status on the PSSA, and special education status. These results expand the research-base on the diagnostic accuracy and appropriateness of use as a screening and decision-making tool for low-performing readers.

4.1.1 Research Questions 1a and 1e Results

Research Questions 1a and 1e

Are the established, author-recommended DIBELS cut-points for each benchmark period in grades one to three accurate predictors of achieving subsequent benchmark goals?

a) Did first graders who achieved NWF and PSF benchmark goals in the Fall achieve NWF, PSF, and DORF benchmark goals in the Winter and Spring?

e) Will the use of ROC analysis result in the establishment of cut-points with greater predictive power including an appropriate balance of sensitivity and specificity?

4.1.1.1 Descriptive Statistics

The descriptive characteristics of first grade 2006 DIBELS performance data are presented in Table 12. Only complete records containing the results of each subtest for each benchmark period were analyzed to facilitate more specific analysis of student progress from Fall to Winter to Spring.

Table 12: Descriptive Statistics- 1st Grade DIBELS 2006

Measure	N	Mean	Range	Std. Deviation
Fall PSF	8595	34.82	0-79	19.21
Fall NWF	8595	27.39	0-142	20.00
Winter PSF	8595	48.19	0-101	17.75
Winter NWF	8595	50.57	0-184	26.12
Winter DORF	8595	30.18	0-213	27.09
Spring PSF	8595	51.07	0-119	15.39
Spring NWF	8595	61.96	0-190	30.45
Spring DORF	8595	47.11	0-184	31.83

Results of the evaluation of assumptions indicated no violations of homoscedastity or multicolinearity.

4.1.1.2 Correlation

A Pearson product moment correlation coefficient matrix was constructed for all DIBELS subtests administered to first grade students in 2006.

Table 13: Correlation Across 1st Grade DIBELS Subtests

Measure	Fall PSF	Fall NWF	Winter PSF	Winter NWF	Winter DORF	Spring PSF	Spring NWF
Fall PSF	_						
Fall NWF	.50	_					
Winter PSF	.55	.33	_				
Winter NWF	.43	.70	.46	_			
Winter DORF	.39	.73	.31	.73	_		
Spring PSF	.41	.25	.61	.34	.22	_	
Spring NWF	.39	.61	.39	.76	.70	.40	_
Spring DORF	.40	.70	.35	.71	.90	.30	.74

All correlation coefficients were significant the $p<.01$ level. However, if considering Spring oral reading fluency as an outcome achievement measure, results suggest that the Fall and Winter NWF subtest results were more strongly related to DORF (.70 & .71 respectively) performance than were Fall and Winter PSF results (.40 & .35 respectively.) Not surprisingly, students' performance on the Winter DORF was most highly linked to the performance on the Spring DORF (.90). The overall significance of the relationships between these subtests should be interpreted cautiously, however, due to the fact that the assessments are administered temporally

close to one another during the school year. Therefore, frequency of testing may have inflated the correlations between measures making the relationships appear stronger than they are.

4.1.1.3 Partial Replication Good et al. (2001) Study

Good et al. (2001) described DIBELS Spring DORF benchmarks as the "anchors" for this prevention-oriented assessment. Spring DORF goals were established first then Winter, then Fall based on analysis of students' slopes of progress across 8 months of school. The same process occurred for establishing the benchmarks for the early literacy subtests (PSF & NWF) in the first grade assessment. Students' progress on those subtests was measured in relation to their end-of-year fluency rates. DIBELS' authors suggest that students who achieve one benchmark goal are likely to achieve subsequent benchmark goals and maintain a pattern of successful achievement. Those students are considered at "low risk" for reading difficulties and are in no need of instructional intervention. Conversely, students who perform significantly below benchmark on one measure are unlikely to achieve subsequent benchmark goals and are considered to be "at risk" for reading difficulties. These students require intensive instructional intervention (Good et al., 2001).

Table 14 explains the consistency of achievement patterns across all DIBELS subtests administered in first grade during the 2006 school year. To parallel the 2001 Oregon study, only the proportion of students in the "low risk" and "at risk" categories were calculated, "some risk" categories were not considered because the likelihood of either successful or unsuccessful future achievement is uncertain. Column 1 lists the measures that are being compared. Columns 2-5 report the percent of students: a) who achieved DIBELS benchmarks in the Fall and Winter and Spring; b) who did not achieve DIBELS benchmarks in the Fall and Winter and Spring; c) who

achieved an earlier but not a subsequent DIBELS benchmark; or d) who did not achieve an earlier benchmark but later experienced success on the DIBELS.

Table 14: Consistency of 1st Grade Students' DIBELS Achievement in 2006

1	2	3	4	5
Measures	**% Students Low Risk/ Low Risk**	**% Students At Risk/ At Risk**	**% Students Low Risk/ At Risk**	**% Students At Risk/ Low Risk**
Fall PSF/ Winter PSF	90%	20%	3%	38%
Fall PSF/ Spring PSF	78%	9%	2%	53%
Fall PSF /Winter DORF	67%	42%	4%	18%
Fall PSF/ Spring DORF	63%	46%	8%	18%
Winter PSF/Spring DORF	84%	22%	10%	9%
Fall NWF/ Winter NWF	67%	46%	4%	8%
Fall NWF/ Spring NWF	76%	27%	3%	12%
Fall NWF/ Winter DORF	77%	42%	3%	13%
Fall NWF/ Spring DORF	72%	49%	3%	14%
Winter NWF/ Spring DORF	68%	45%	2%	8%

Column 2: Low Risk Who Stayed Low Risk

According to the DIBELS authors, achieving benchmark proficiency in the Fall is supposed to predict benchmark proficiency on later assessments. Five thousand sixty-six first graders were "low risk" on the PSF subtest in the Fall. Ninety percent of those students maintained "low risk" status in the Winter (n=4,558); 78% maintained that successful status in the Spring (n=3,951). Similar results occurred for the NWF subtests. When comparing Fall to Winter to Spring achievement, 67% of the 4,802 students considered "low risk" early in the school year were also considered "low risk" by mid-year whereas 76% continued to achieve benchmark goals in the Spring. "Low risk" status on the PSF and NWF subtests in the Fall also predicted "low risk" status on the DORF in the Winter and Spring. Sixty-seven percent of the students who achieved Fall PSF goals also achieved Winter DORF goals and 63% achieved Spring DORF goals (n=3,394 & 3,191 respectively). Furthermore, 77% of students meeting NWF goals in the fall met DORF goals in the winter and 72% of those students met DORF goals in the Spring (n=3,697 & 3,457 respectively).

Column 3: At Risk Who Stayed At Risk

When examining the results for the 1,295 students who did not meet benchmark goals in the Fall (students "at risk") we see that 20% or 259 of those low performing students continued to have difficulty in the Winter and 117 students (9%) performed below benchmark in the Spring on the PSF subtest. Similar patterns occurred on the NWF subtest. Nearly half (46%) of the 2,147 students who struggled on the NWF subtest in the Fall were also "at risk" in the Winter and 1/3 (27%) continued to be at risk in the Spring on that subtest. DORF outcomes show that 42% of the students who performed below benchmark on the Fall PSF subtest scored below benchmark on the Winter DORF and 46% scored below benchmark on the Spring DORF (n=544

& 596 respectively). Likewise, 2,016 students (42%) who performed below benchmark on the Fall NWF subtest scored below benchmark on the Winter DORF. Almost half of those students at risk on the NWF in the Fall (49%) also scored below benchmark on the Spring DORF (n=2,352).

Column 4: Low Risk Who Became At Risk

While the patterns of consistent achievement were interesting, examining inconsistent growth and movement of students between performance categories sheds more light on the utility of the DIBELS benchmark goals. These analyses focus on the students whose early DIBELS scores did not predict future DIBELS scores. Fall to Spring outcomes on the PSF subtest reveal that while the proportion of students are low, some variability in achievement did occur for students originally meeting performance standards. Specifically, 2% of the 5,066 students meeting PSF benchmark goals in the Fall were considered to be in need of more intensive instructional interventions by the Spring PSF administration (n=102). NWF achievement parallels that pattern with only 3% of the 4,802 "low risk" students moving to "at risk" by Spring (n=162). When using Spring DORF scores as the yardstick for measuring students' progress throughout the year, the data show that 8% of the students who met PSF benchmark goals in the Fall failed to meet grade-level oral reading fluency standards in the Spring (n=384). Likewise, 3% of the students meeting or exceeding Fall NWF goals performed well below expectations on the Spring DORF subtest (n=154). DIBELS did not predict in the way that it was supposed to for those students.

Column 5: At Risk Who Became Low Risk

The proportions of students moving from "at risk" to "low risk" also raise some important questions. The data indicate that 38% of the 1,295 students achieving well below benchmark in the Fall on the PSF subtest achieved goals in the Winter whereas an astounding 53% achieved benchmark in the Spring. That means that 1,137 students flagged for reading failure on the basis of their September PSF scores met or exceeded PSF goals in the Spring. On the NWF subtest, 12% of the students struggling with alphabetic principal tasks in the Fall, mastered them by Spring (n=257). Additionally, 18% of students considered "at risk" on the PSF in the Fall were not "at risk" on the DORF in the Winter and again met Spring DORF expectations (n=386). Among students scoring "at risk" on the NWF in the Fall, 13% achieved adequate levels of fluency on the Winter DORF compared to 14% on the Spring DORF (n=279 & 300 respectively). Fall DIBELS benchmark achievement levels were not accurate predictors of future performance for those students.

One explanation for the changes in risk status might be the positive impact of the intensive instructional interventions delivered to these students based on their low Fall performance within the Reading First framework. Another may be that the author recommended cut-points for identifying students as "at risk" for reading failure at the Fall testing time are too sensitive resulting in over-identification of students in need of support. The possibility of the latter leads to an analysis of the diagnostic accuracy of DIBELS achievement thresholds for the first grade subtests.

4.1.1.4 Diagnostic Accuracy Analysis

Analyses of the diagnostic accuracy of the DIBELS subtests for 2006 first grade achievement was conducted simultaneously as an additional feature of the ROC analysis. The analyses generated the ratio of true positive/false positive and true negative/false negative decisions of students at "low risk" and "at risk" based on the author-recommended Fall cut-points for the PSF and NWF subtests. The analysis examines two possibilities—the presence of the tested condition or absence of the tested condition. In reference to the DIBELS performance descriptors, that means when testing for "at risk" status (i.e. students predicted to be poor readers) true positive decisions result when those students actually become poor readers. False positive decisions results when those "at risk" students do not become poor readers. Furthermore, when testing for "low risk" status (i.e. students predicted to be good readers) true negative decisions result when those students actually become good readers. False negative decisions result when those "low risk" students do not become good readers.

The following 2x2 matrix illustrates the population of students discussed in this section. The Spring DORF benchmark levels were used as outcome measures for these analyses because DIBELS authors use end-of-year fluency rates as indicators of overall reading skill for students in grades 1-3.

		Outcome Measure	
		Spring DIBELS DORF	
Prediction Measure		Low Risk	Some Risk or At Risk
Fall DIBELS subtests	Low Risk	*A*	*B*
	At Risk	*D*	*C*

A: students who achieved "low risk" status in the Fall who also achieved "low risk" status in the Spring.

B: students who achieved "low risk" status in the Fall who achieved "some risk" or "at risk" status in the Spring.

D: students who achieved "at risk" status in the Fall who achieved "at risk" status in the Spring.

C: students who achieved "at risk" status in the Fall who achieved "low risk" or "some risk" status in the Spring.

Figure 2: Decision Matrix for Fall to Spring DIBELS Predictions

Table 15: Diagnostic Accuracy of First Grade DIBELS PSF Benchmark to Predict "Low Risk" Status on Spring DORF

Cut-point	True Negative Decisions *(A)*	False Negative Decisions *(B)*
35 cspm	63%	15%

Five thousand sixty-six first grade students fell into the "low risk" performance category for the Fall PSF subtest. The results show that 63% of those "low risk" students were accurately identified when using a cut-point of 35 correct sounds per minute on the Fall PSF test to predict Spring risk status the DORF. Diagnostically speaking, those 3,191 first graders fell into box *A* in

Figure 2. They were truly "low risk" and continued to succeed in reading as predicted by the Fall PSF DIBELS test results. Conversely, 15% of "low risk" students were misidentified using the same cut-point. Those students fell into box *B* in Figure 2. Essentially, 760 students actually in need of additional reading support (i.e. they ended the year "at risk" on DORF) did not receive it.

Table 16: Diagnostic Accuracy of First Grade DIBELS PSF Benchmark to Predict "At Risk" Status on Spring DORF

Cut-point	True Positive Decisions *(C)*	False Positive Decisions *(D)*
10 cspm	19%	26%

One thousand two hundred ninety-five students were categorized as "at risk" on the Fall PSF subtest. Nineteen percent were accurately identified using a cut-point of 10 correct sounds per minute on the Fall PSF test to predict Spring outcomes (box *C* in Figure 2). That is to say that 246 first graders flagged for reading intervention truly needed reading intervention. Inaccurate identification occurred for 26% of the participating first graders who were initially labeled "at risk;" by Spring these students were no longer considered "at risk" based on DORF scores (box *D* in Figure 2). Given the prevention framework of Reading First, those students were candidates for strategic or intensive reading instruction including more instructional time, more focused work, and more opportunities for small-group and one-on-one work with a teacher. In reality however, those 336 students my not have needed intervention. The diagnostic results will be explored further when examining the PSF ROC Curve and score continuum for "at risk" status.

Table 17: Diagnostic Accuracy of First Grade DIBELS NWF Benchmark to Predict "Low Risk" Status on Spring DORF

Cut-point	True Negative Decisions *(A)*	False Negative Decisions *(B)*
24 cspm	74%	17%

Four thousand eight hundred two students achieved benchmark on the Fall NWF subtest. The results show that 74% (n=3,553) of those "low risk" students were accurately identified using a cut-point of 24 correct sounds per minute on the Fall NWF subtest test to predict Spring outcomes (box *A* Figure 2). Conversely, 17% (n=816) of "low risk" students were identified incorrectly; the DIBELS results were not accurate indicators of future successful benchmark achievement for those students.

Table 18: Diagnostic Accuracy of First Grade DIBELS NWF Benchmark to Predict "At Risk" Status on Spring DORF

Cut-point	True Positive Decisions *(C)*	False Positive Decisions *(D)*
13 cspm	85%	14%

The analysis indicated that 85% of the 2,147 "at risk" students were accurately identified using a cut-point of 13 correct sounds per minute on the Fall NWF test to predict Spring outcomes. Those students exhibited the consistent achievement patterns explained in box *C* of Figure 2.

Inaccurate identification occurred for 14% (n=300) of the participating first graders. Those students fell into box *D*.

4.1.1.5 ROC Curve Analysis

To further explore the utility of the author recommended cut-points with respect to sensitivity and specificity, the area and coordinates of 4 different ROC Curves were examined. Decision criterion levels for selecting accurate cut-points were based on Swets (1988) recommendation that the area under the curve should be greater than .75 to achieve an appropriate balance of sensitivity and specificity.

Table 19: ROC Curve Descriptions for 1st Grade Fall to Spring Predictions

Curve	Outcome	Predictor Measure	Outcome Measure
1	Successful (Low Risk)	Fall PSF	Spring DORF
2	Unsuccessful (At Risk)	Fall PSF	Spring DORF
3	Successful (Low Risk)	Fall NWF	Spring DORF
4	Unsuccessful (At Risk)	Fall NWF	Spring DORF

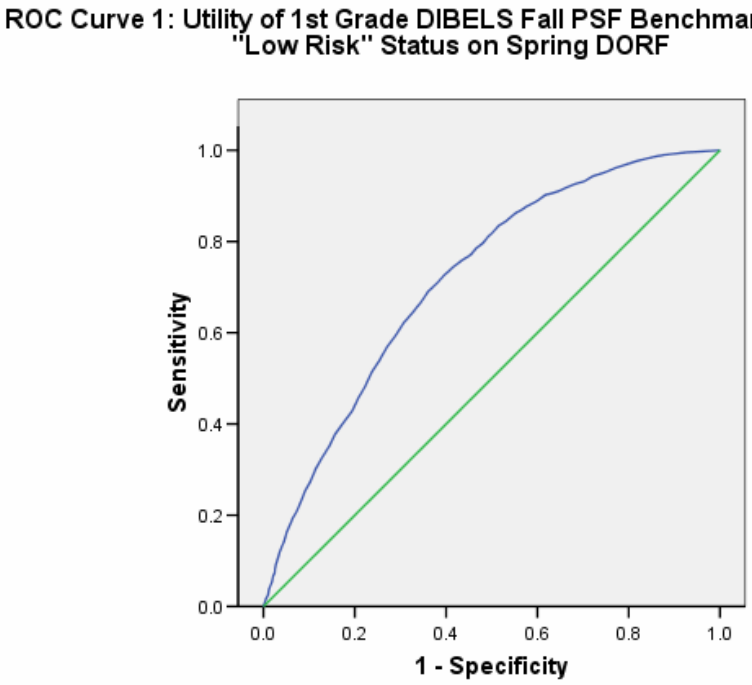

Figure 3: ROC Curve 1

Table 20: ROC Curve 1 Summary Table: 1st Grade 2006

	DIBELS Cut-point	**True Negative Decisions**	**False Negative Decisions**
	35 cspm	63%	15%
Area Under the Curve	**Recommended Cut-Point to achieve .75**	**Resulting True Negative Decisions**	**Resulting False Negative Decisions**
.71	30 cspm	80%	17%

* Comparison of Sensitivity/Specificity Balance of the Author Recommended Cut-point for Low Risk Status on the Fall PSF Subtest to the Sensitivity/Specificity Balance of a Newly Recommended Cut-Point for Low Risk Status on the Fall PSF Subtest

The ROC curve produced using the currently established cut point of 35 correct sounds or phonemes produced in one minute for students Fall performance on the PSF subtest did not meet acceptability requirements (area=.71) indicating an inappropriate balance of sensitivity and specificity. Analysis results suggest that to increase the area under the curve to a standard of .75, the cut point should be decreased by 5 words thereby increasing the accurate decisions for "low risk" status to 80%. However, this change would result in an increase in the number of false negative decisions (17%) meaning that even though a greater number of students would be accurately identified as "low risk," more students would also be inaccurately identified.

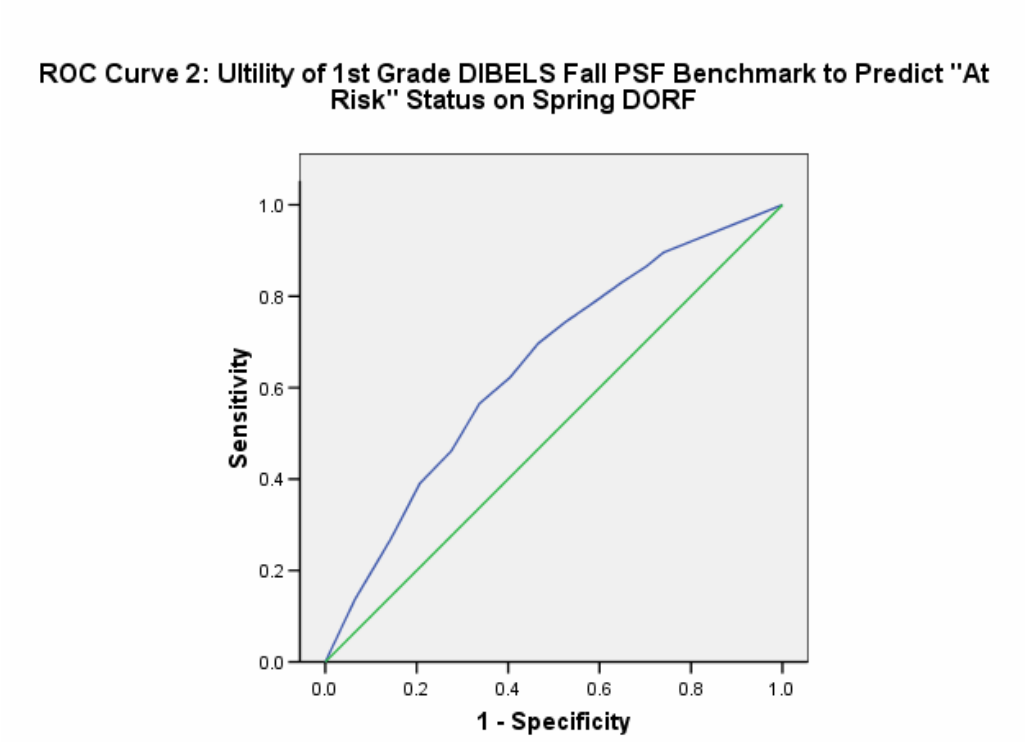

Figure 4: ROC Curve 2

Table 21: ROC Curve 2 Summary Table: 1st Grade 2006

	DIBELS Cut-point	True Positive Decisions	False Positive Decisions
	10 cspm	19%	26%
Area Under the Curve	**Recommended Cut-Point to achieve .75**	**Resulting True Positive Decisions**	**Resulting False Positive Decisions**
.64	8 cspm	60%	29%

* Comparison of Sensitivity/Specificity Balance of the Author Recommended Cut-point for At Risk Status on the Fall PSF Subtest to the Sensitivity/Specificity Balance of a Newly Recommended Cut-Point for At Risk Status on the Fall PSF Subtest

The ROC curve produced using the currently established cut point of at least 10 correct sounds produced in one minute for students labeled "at-risk" on the PSF subtest did not meet acceptability requirements (area=.64) indicating an inappropriate balance of sensitivity and specificity in predicting spring outcomes on the DORF. Analysis results suggest that to increase the area under the curve to a standard of .75 the cut point should be set at 8 cspm. The lowering of this threshold would result in 60% of students correctly being identified "at risk". However, the new cut point would also result in a 3 percent increase in inaccuracies.

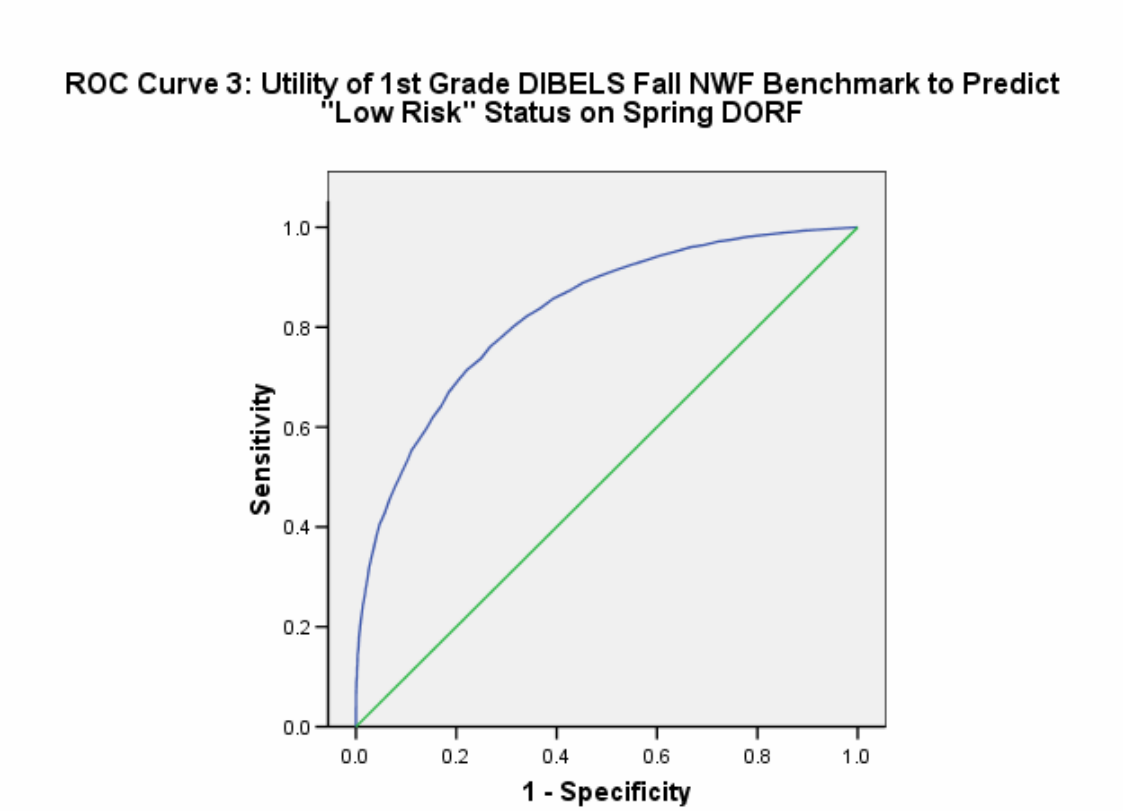

Figure 5: ROC Curve 3

Table 22: ROC Curve 3 Summary Table: 1st Grade 2006

	DIBELS Cut-point	**True Negative Decisions**	**False Negative Decisions**
	24 cspm	74%	17%
Area Under the Curve	**Recommended Cut-Point to achieve .75**	**Resulting True Negative Decisions**	**Resulting False Negative Decisions**
.83	NA	NA	NA

* Comparison of Sensitivity/Specificity Balance of the Author Recommended Cut-point for Low Risk Status on the Fall NWF Subtest to the Sensitivity/Specificity Balance of a Newly Recommended Cut-Point for Low Risk Status on the Fall PSF Subtest

The ROC curve produced using the currently established cut point of 24 correct sounds produced in one minute for students Fall performance on the NWF exceeded acceptability requirements (area=.83) indicating an appropriate balance of sensitivity and specificity. There is no need to establish new cut-points. The proportions of students identified both accurately and inaccurately by the Fall NWF subtest fall within diagnostic acceptability guidelines.

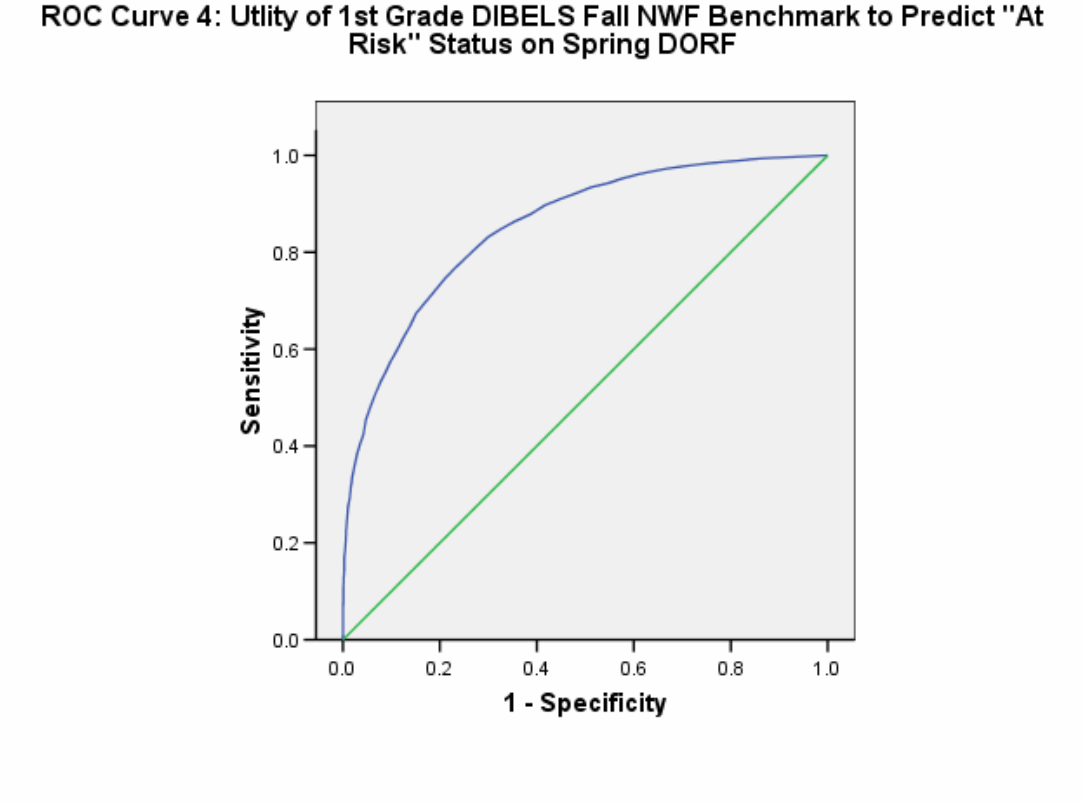

Figure 6: ROC Curve 4

Table 23: ROC Curve 4 Summary Table: 1st Grade DIBELS 2006

	DIBELS Cut-point	True Positive Decisions	False Positive Decisions
	13 cspm	85%	14%
Area Under the Curve	**Recommended Cut-Point to achieve .75**	**Resulting True Positive Decisions**	**Resulting False Positive Decisions**
.85	NA	NA	NA

* Comparison of Sensitivity/Specificity Balance of the Author Recommended Cut-point for At Risk Status on the Fall NWF Subtest to the Sensitivity/Specificity Balance of a Newly Recommended Cut-Point for At Risk Status on the Fall PSF Subtest

Like the curve 3, ROC curve 4 produced using the currently established cut point of at least 13 correct sounds produced in one minute for students labeled "at-risk" on the NWF exceeded the standard of .75 (area= .85) indicating an appropriate balance of sensitivity and specificity in predicting spring outcomes on the DORF. Overall, the diagnostic utility of the NWF subtest is acceptable.

4.1.2 Research Questions 1b and 1e Results

Research Questions 1b and 1e

Are the established, author-recommended DIBELS cut-points for each benchmark period in grades one to three accurate predictors of achieving subsequent benchmark goals?

b) Did second graders who achieved DORF benchmark goals in the Fall achieve DORF benchmark goals in the Winter and subsequent DORF benchmark goals in the Spring?

e) Will the use of ROC analysis result in the establishment of cut-points with greater predictive power including an appropriate balance of sensitivity and specificity?

4.1.2.1 Descriptive Statistics

The descriptive characteristics of second grade 2006 DIBELS performance data are presented in Table 24. Only complete records containing the results of each subtest for each benchmark period were analyzed to facilitate more specific analysis of student progress from Fall to Winter to Spring.

Table 24: Descriptive Statistics: 2nd Grade DIBELS 2006

Measure	N	Mean	Range	Std. Deviation
Fall DORF	7925	44.9	0-209	27.93
Winter DORF	7925	71.0	0-216	35.12
Spring DORF	7925	83.0	0-235	35.54

Results of the evaluation of assumptions indicated no violations of homoscedacity or multicolinearity. The increase in the average number of words read correctly at each benchmark period suggests that students' fluency skills strengthened throughout the year. The variability in scores also increased.

4.1.2.2 Correlation

A Pearson product moment correlation matrix was calculated for all DIBELS subtests administered to second graders in 2006.

Table 25: Correlation Across 2nd Grade DIBELS Subtests

Measure	Fall DORF	Winter DORF
Fall DORF	_	
Winter DORF	.90	_
Spring DORF	.86	.93

All correlations were significant at the $p<.01$ level. Similar to the first grade results, the high correlations between each subtest is expected due to the timing of administrations. The significant relationships should be interpreted cautiously due to the fact that the assessments were administered temporally close to one another during the school year. It is also important to note that high correlations are expected between these DORF subtests because they are essentially the same measure (i.e. one minute calculations of students' ability to read grade-level passages aloud with speed and accuracy).

4.1.2.3 Partial Replication Good et al. (2001) Study

Table 26 explains the linkages between second graders' earlier and subsequent achievement on the 2006 DIBELS. As discussed earlier, comparisons were made for students performing at "low risk" and "at risk" only, students "some risk" were not included. Column 1 indicates the measures of comparison. Columns 2-5 report the percent of students: a) who achieved DORF benchmarks in the Fall and Winter and Spring; b) who did not achieve DORF

benchmarks in the Fall and Winter and Spring; c) who achieved an earlier but not a subsequent DORF benchmark; or d) who did not achieve an earlier benchmark but later experienced success on the DORF.

Table 26: Consistency of 2nd Grade Students' DIBELS Achievement in 2006

1	2	3	4	5
Measures	**% Students Low Risk/ Low Risk**	**% Students At Risk/ At Risk**	**% Students Low Risk/ At Risk**	**% Students At Risk/ Low Risk**
Fall DORF/ Winter DORF	89%	78%	3%	6%
Fall DORF/ Spring DORF	77%	75%	2%	7%
Winter DORF/ Spring DORF	76%	82%	2%	2%

Column 2: Low Risk Who Stayed Low Risk

Three thousand eight hundred ninety-five students reached or exceeded DIBELS second grade DORF benchmark goals in the Fall. Three thousand four hundred ninety-two (89%) of those students also met DORF goals in the Winter and 77% continued to perform well in the Spring (n=3,004). Winter assessment results showed that 4,635 students read proficiently enough to be considered "low risk." Seventy-six percent of those fluent readers maintained "low risk" status from Winter to Spring (n=3535).

Column 3: At Risk Who Stayed At Risk

Fluency performance for the "at risk" subgroup was also relatively consistent. Two thousand one hundred sixty-nine students achieved well below grade-level DIBELS standards on the Fall DORF. 78% of students in that "at risk" category scored "at risk" again in the Winter (n= 1,692). Only 57 students changed risk status by Spring, while the other 75% remained "at risk." Similarly, 82% of the 2,067 students who did not meet 2nd grade performance expectations in the Winter continued to struggle in the Spring (n=2,520). Those figures suggest that overall DIBELS estimates of students' oral reading fluency rates were relatively consistent throughout 2nd grade.

Column 4: Low Risk Who Became At Risk

The movement of students between benchmark categories was not as frequent for second grade as it was for first. For example, only 3% of the 3,895 high performing students on the Fall DORF were "at risk" by Winter (n=134). Eighty-eight (2%) of the students who exceeded Fall DORF goals of reading at least 44 wpm read below 77 words per minute on the Spring DORF measure, thereby falling into the "at risk" category. Similarly, 2% of the students who were "low risk" on the Winter DORF, dropped to "at risk" on the Spring assessment.

Column 5: At Risk Who Became Low Risk

Changes in the proportion of students from "at risk" to "low risk" were also interesting. 2,169 students read less than 26 words per minute on the Fall DORF measure, which placed them into the "at risk" benchmark category. 6% of those students increased their fluency rate by at least 43 words to achieve "low risk" status on the Winter test (n=139). Rate increases were even more dramatic for 2% of that Fall "at risk" subgroup whose Spring DORF scores exceeded

90 word per minute placing them at "low risk." Winter to Spring comparisons showed that 20 of the 2,500 students "at risk" in January met benchmark goals by May.

4.1.2.4 Diagnostic Accuracy Analysis

The framework for the discussion of the diagnostic accuracy results for second grade students in 2006 mirrors the aforementioned discussion of first grade results. Refer back to Figure 2 for further clarification.

The percentages of students correctly or incorrectly identified as "low risk" in the Spring based on Fall DORF performance are listed in the table below.

Table 27: Diagnostic Accuracy of Second Grade DIBELS Fall DORF Benchmark to Predict "Low Risk" Status on Spring DORF

Cut-point	True Negative Decisions *(A)*	False Negatives Decisions *(B)*
44 cwpm	84%	11%

The results show that 84% of 3,895 students considered "low risk" on the Fall DORF were accurately identified by the Fall DORF using a cut-point of 44 correct words per minute. The prediction of "low risk" status on the Spring DORF was strong for those students falling into box *A* in Figure 2. However, 11% of "low risk" students were misidentified using the same cut-point; meaning, Fall performance was not predictive of Spring performance for 428 second graders.

Table 28: Diagnostic Accuracy of Second Grade DIBELS Fall DORF Benchmark to Predict "At Risk" Status on Spring DORF

Cut-point	True Positive Decisions *(C)*	False Positive Decisions *(D)*
26 cwpm	88%	10%

The analysis indicated that 88% of "at risk" students were accurately identified using a cut-point of 26 words correct per minute on the Fall PSF subtest when predicting Spring outcomes. Therefore, 1,909 students labeled in need of intensive reading instruction according to Fall DIBELS benchmarks were truly "at risk". Inaccurate identification occurred for only 10% of "at risk" second graders studied. Those 216 students were false positives and were not truly "at risk". Intensive reading intervention was unnecessary for those students as they would likely be in the population of students who later met Winter and/or Spring DORF goals.

4.1.2.5 ROC Curve Analysis

To determine whether the aforementioned numbers of true negative and false negative vs. true positive and fall positive identifications are statistically appropriate, 2 different ROC Curves were examined. Decision criterion levels for selecting accurate cut-points were based on Swets (1988) recommendation that the area under the curve should be greater than .75 to achieve a desirable balance of sensitivity and specificity.

Table 29: ROC Curve Descriptions for 2nd Grade Fall to Spring Predictions

Curve	Outcome	Predictor Measure	Outcome Measure
5	Successful (Low Risk)	Fall DORF	Spring DORF
6	Unsuccessful (At Risk)	Fall DORF	Spring DORF

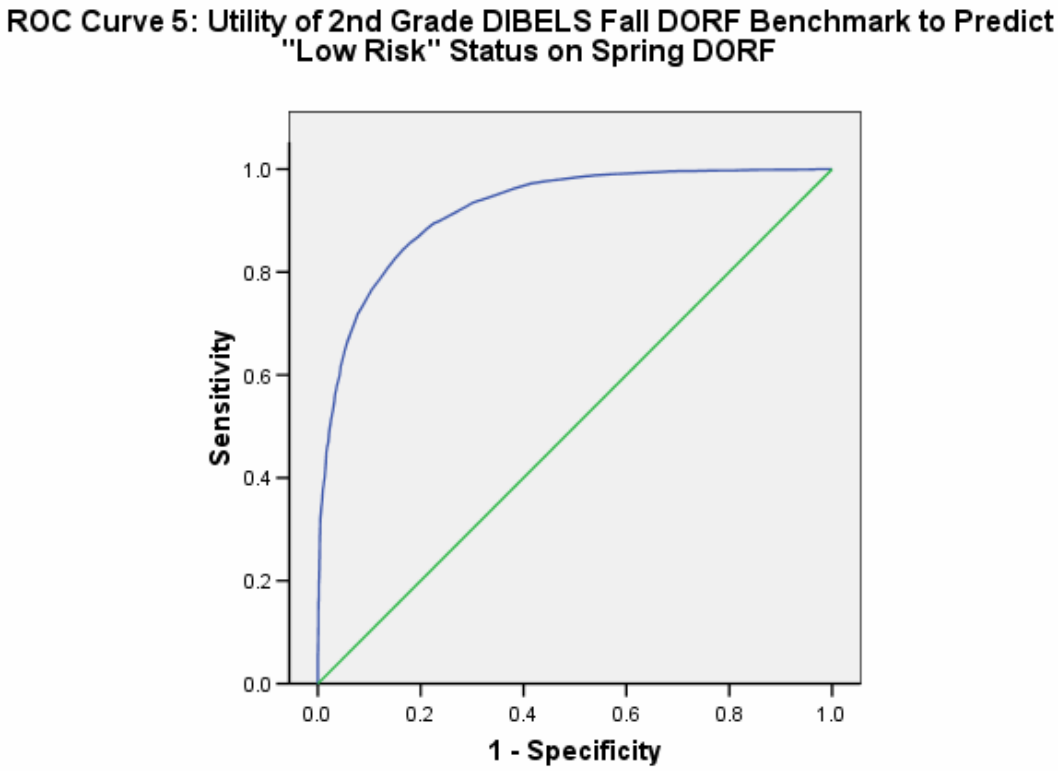

Figure 7: ROC Curve 5

Table 30: ROC Curve 5 Summary Table: 2nd Grade DIBELS 2006

	DIBELS Cut-point	**True Negative Decisions**	**False Negative Decisions**
	44 cwpm	84%	11%
Area Under the Curve	**Recommended Cut-Point to achieve .75**	**Resulting True Negative Decisions**	**Resulting False Negative Decisions**
.92	NA	NA	NA

***** Comparison of Sensitivity/Specificity Balance of the Author Recommended Cut-point for Low Risk Status on the Fall DORF Subtest to the Sensitivity/Specificity Balance of a Newly Recommended Cut-Point for Low Risk Status on the Fall DORF Subtest

The ROC curve produced using the currently established DORF cut point of 44 correct words produced in one minute for students Fall fluency rates exceeded acceptability requirements (area=.92) indicating an appropriate balance of sensitivity and specificity. There was no need to establish new cut-points.

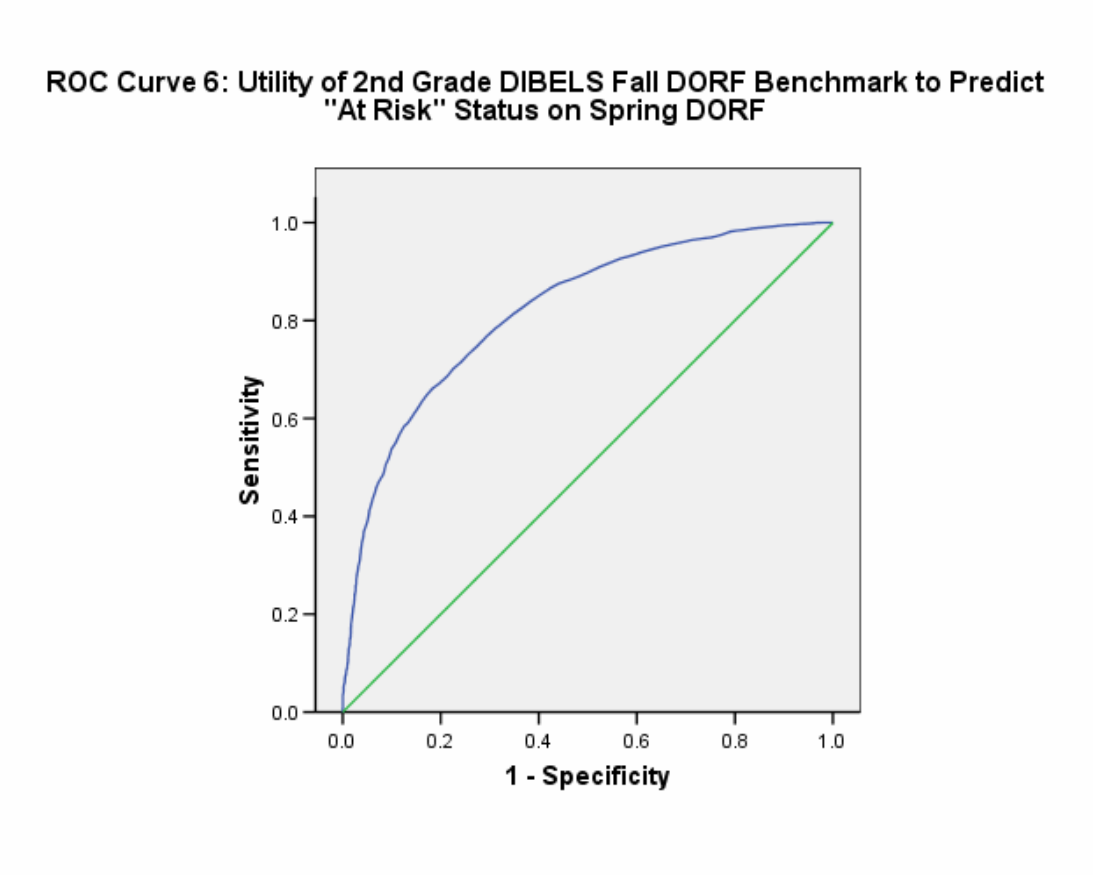

Figure 8: ROC Curve 6

Table 31: ROC Curve 6 Summary Table: 2nd Grade DIBELS 2006

	DIBELS Cut-point	True Positive Decisions	False Positive Decisions
	26 cwpm	88%	10%
Area Under the Curve	**Recommended Cut-Point to achieve .75**	**Resulting True Positive Decisions**	**Resulting False Positive Decisions**
.82	NA	NA	NA

* Comparison of Sensitivity/Specificity Balance of the Author Recommended Cut-point for At Risk Status on the Fall DORF Subtest to the Sensitivity/Specificity Balance of a Newly Recommended Cut-Point for At Risk Status on the Fall DORF Subtest

The ROC curve for a Fall DORF cut point of 26 cwpm shows that it was an accurate predictor of "at risk" performance in the Spring overall. Although 10% of the decisions were found to be false positives (i.e. not truly "at risk"), the diagnostic accuracy and decision making utility estimates fall within acceptable parameters. The measure has an acceptable balance of sensitivity and specificity.

4.1.3 Research Questions 1c and 1e Results

Research Questions 1c and 1e

Are the established, author-recommended DIBELS cut-points for each benchmark period in grades one to three accurate predictors of achieving subsequent benchmark goals?

c) Did third graders who achieved DORF benchmark goals in the Fall achieve DORF benchmark goals in the Winter and subsequent DORF benchmark goals in the Spring?

e) Will the use of ROC analysis result in the establishment of cut-points with greater predictive power including an appropriate balance of sensitivity and specificity?

4.1.3.1 Descriptive Statistics

The descriptive characteristics of third grade 2006 DIBELS performance data are presented in Table 32. Only complete records containing the results of each subtest for each benchmark period were analyzed to facilitate more specific analysis of student progress from Fall to Winter to Spring.

Table 32: Descriptive Statistics: 3rd Grade DIBELS 2006

Measure	N	Mean	Range	Std. Deviation
Fall DORF	8317	69.13	0-205	31.68
Winter DORF	8317	84.75	0-242	35.22
Spring DORF	8317	98.43	0-253	35.86

Results of the evaluation of assumptions indicated no violations of homoscedacity or multicolinearity. The average number of words read correctly per minute increased by twenty-nine words from Fall to Spring. The variability of scores also increased.

4.1.3.2 Correlation

A Pearson product moment correlation matrix was constructed for the 2006 third grade DIEBLS results.

Table 33: Correlation Across 3rd Grade DIBELS Subtests

Measure	Fall DORF	Winter DORF
Fall DORF	_	
Winter DORF	.91	_
Spring DORF	.88	.92

The reported statistics suggest that students' scores on the DORF during each benchmark period in third grade were highly related. Similar to the second grade results, this relationship was

strongest between Fall to Winter subtests (.91) and Winter to Spring subtests (.92). Again, the high correlations between each subtests is expected. The linkages should be interpreted cautiously due to the fact that the assessments were administered within an eight-month period. The correlations are probably also high because the assessment measures were the same (i.e. measures of oral reading fluency).

4.1.3.3 Partial Replication Good et al. (2001) Study

In partial replication of the Good et al. (2001) study, the relationship between earlier and subsequent achievement were examined by comparing 3rd grade students' benchmark across each DORF benchmark administration in the Fall, Winter, and Spring of 2006. Only students labeled "low risk" and "at risk" were part of the analysis, students labeled "some risk" were not. In Table 34, columns 2 and 3 report the percentage of students maintaining consistent patterns of achievement on designated DIBELS subtests. Columns 4 and 5 include the percentage of students demonstrating inconsistent performance on designated DIBELS subtests from one benchmark period to the next.

Table 34: Consistency of 3rd Grade Students' DIBELS Achievement in 2006

1	2	3	4	5
Measures	**% Students Low Risk/ Low Risk**	**% Students At Risk/ At Risk**	**% Students Low Risk/ At Risk**	**% Students At Risk/ Low Risk**
Fall DORF/ Winter DORF	85%	76%	2%	3%
Fall DORF/ Spring DORF	76%	64%	3%	4%
Winter DORF/ Spring DORF	79%	71%	3%	5%

Column 2: Low Risk Who Stayed Low Risk

When examining the consistency of benchmark classification we see that 85% of the 3,216 students who met DORF benchmark goals in the Fall also met them in the Winter, whereas; 76% also met them in the Spring (n=2,729 & 2,441 respectively). Furthermore, there were 3,433 third grade students who read with enough fluency in the Winter to achieve "low risk" status on the January DORF subtest. Seventy-nine percent of those "low risk" students maintained proficiency levels on the Spring DORF (n=2,728). These statistics suggest that the majority of students who were strong readers at the beginning of third grade were strong readers at the end of third grade. DIBELS' prediction of future performance for those students held true for each assessment period.

Column 3: At Risk Who Stayed At Risk

Over three fourths of the 2,692 students who were "at risk" in the Fall, were also "at risk" in the Winter (76%). Sixty-four percent of the students who did not achieve benchmark in Fall did not achieve benchmark in the Spring. Finally, 71% of the 2,658students flagged as "at risk" based on poor Winter fluency scores, were also flagged as "at risk" in the Spring (n=1,878). Again, DIBELS performance classifications were consistent for those students--students who were poor readers at the beginning of third grade were poor readers at the end of third grade.

Column 4: Low Risk Who Became At Risk

These predictable patterns of achievement were further substantiated by the lack of movement between benchmark categories. Only 2% of students who were successful in reaching optimal fluency rates the Fall were unsuccessful in the Winter (i.e. 76 out of 3,216 students). Three percent of that same population performed poorly on the Spring DORF meaning that 14 more students changed from "low risk" status to "at risk" status from the September to May testing. Similarly, 90 (3%) of the students who reached Winter DORF standards performed well below standards and dropped to "at risk" status by the end of the year.

Column 5: At Risk Who Became Low Risk

The patterns of change were similar for students who moved from "at risk" status to "low risk" status across the three DORF administrations in third grade. End-of-year scores showed that 4% of the students who were formerly "at risk" in the Fall became "low risk" by Spring (i.e. 90 out of 3216 students). Two thousand six hundred fifty-eight students met Winter benchmark goals by reading over 92 wpm but 142 (5%) of those students read below the "at risk" cut-point of 80 wpm in the Spring.

While the aforementioned percentages suggest that Fall DIBELS DORF cut-points for third grade were accurate predictors of Spring DORF performance for the majority of the students tested, further analysis of diagnostic accuracy results and ROC Curves will provide more statistical information about their decision-making utility. These analyses help to substantiate or challenge the appropriateness of instructional grouping decisions for students based on DIBELS instructional recommendations for varying levels of reading intervention linked to risk status.

4.1.3.4 Diagnostic Accuracy Analysis

The framework for the discussion of the diagnostic accuracy results for third grade students in 2006 mirrors the aforementioned discussion of first and second grade results.

The percentages of students correctly or incorrectly identified as "low risk" in the Spring based on Fall DORF performance are listed in the table below.

Table 35: Diagnostic Accuracy of Third Grade DIBELS Fall DORF Benchmark to Predict "Low Risk" Status on Spring DORF

Cut-point	True Negative Decisions *(A)*	False Negative Decisions *(B)*
77 cwpm	85%	14%

The results show that 450 out of 3,216 third graders were misidentified as "low risk" using the Fall DORF cut-point and 85% were accurately identified (box *A* Figure 2). So, one seventh of the students categorized as "low risk" on DIBELS may have benefited from additional reading

support and more intensive instruction but probably did not receive that intervention given their high DIBELS scores. DIBELS DORF scores were not accurate predictors of future achievement for those students (box *B* Figure 2).

Table 36: Diagnostic Accuracy of Third Grade DIBELS Fall DORF Benchmark to Predict "At Risk" Status on Spring DORF

Cut-point	True Positive Decisions *(C)*	False Positive Decisions *(D)*
53 cwpm	79%	20%

The analysis indicated that 79% of "at risk" students were accurately identified using a cut-point of 53 words correct per minute on the Fall DORF subtest when predicting Spring outcomes. Those students fell into box *C* in Figure 2 and experienced predictable patterns of low achievement on the DIBELS DORF from the beginning to the end of the year. Therefore, 2,127 third graders who were flagged for needing intensive reading instruction were truly "at risk". Inaccurate identification occurred for 538 of the third graders falling into the "at risk" DIBELS category (box *D* Figure 2). Intensive reading intervention may have been unnecessary for these students.

4.1.3.5 ROC Curve Analysis

While the percentages of inaccuracies appears small, in-depth analysis based on two different ROC curves generated for the Fall DORF cut-points helped to determine the appropriate balance of true positives/true negative vs. false positive/false negative decisions in

third grade. Decision criterion levels for selecting accurate cut-points were based on Swets (1988) recommendation that the area under the curve should be greater than .75 to achieve an appropriate balance of sensitivity and specificity.

Table 37: ROC Curve Descriptions for 3rd Grade Fall to Spring Predictions

Curve	Outcome	Predictor Measure	Outcome Measure
7	Successful (Low Risk)	Fall DORF	Spring DORF
8	Unsuccessful (At Risk)	Fall DORF	Spring DORF

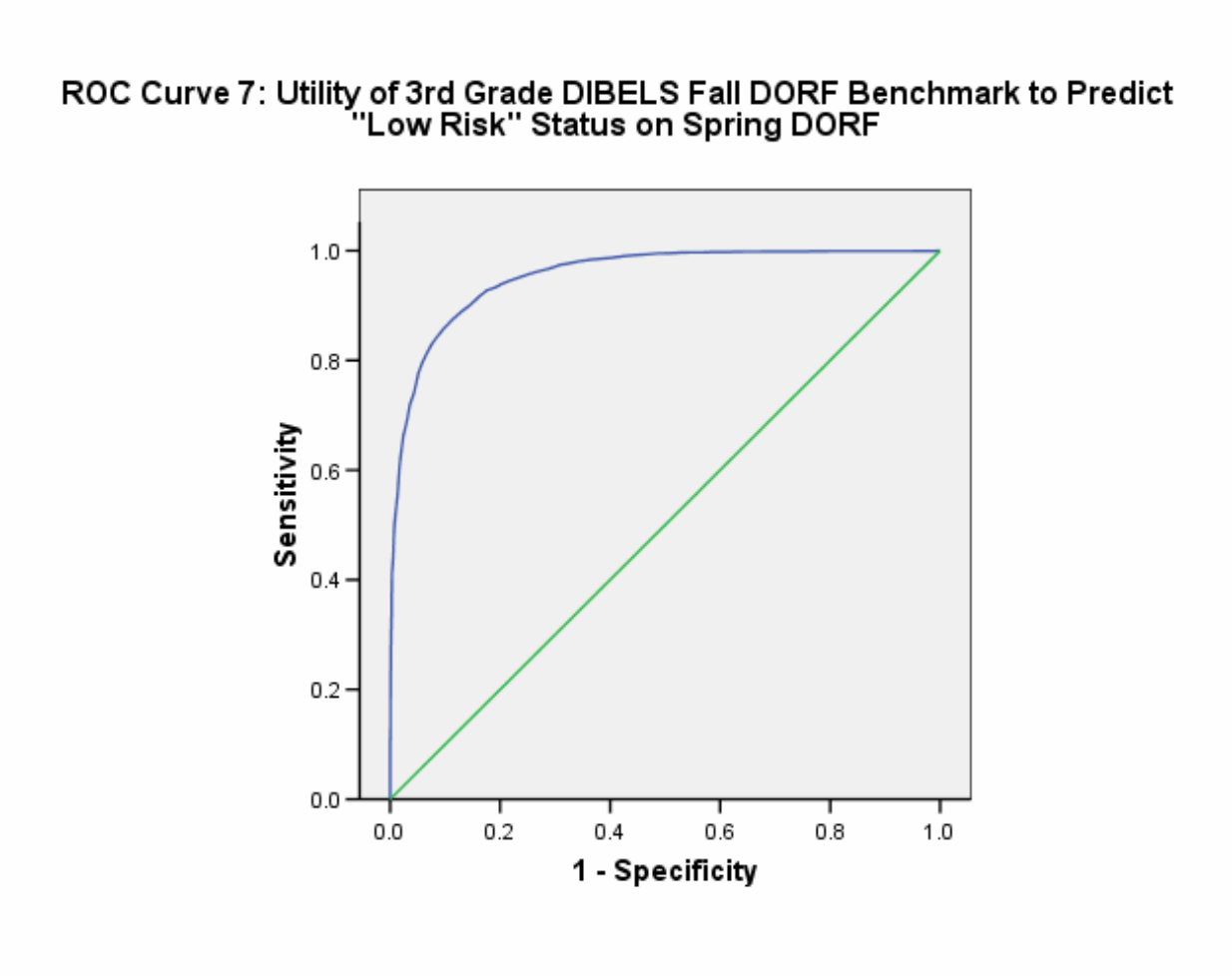

Figure 9: ROC Curve 7

Table 38: ROC Curve 7 Summary Table: 3rd Grade DIBELS 2006

	DIBELS Cut-point	True Positive Decisions	False Positive Decisions
	77 cwpm	85%	14%
Area Under the Curve	**Recommended Cut-Point to achieve .75**	**Resulting True Positive Decisions**	**Resulting False Positive Decisions**
.96	NA	NA	NA

* Comparison of Sensitivity/Specificity Balance of the Author Recommended Cut-point for Low Risk Status on the Fall DORF Subtest to the Sensitivity/Specificity Balance of a Newly Recommended Cut-Point for Low Risk Status on the Fall DORF Subtest

The ROC curve produced using the currently established Fall third grade DORF cut point of 77 correct words per minute substantially exceeded acceptability requirements (area=.96) indicating an appropriate balance of sensitivity and specificity. There was no need to establish new cut-points.

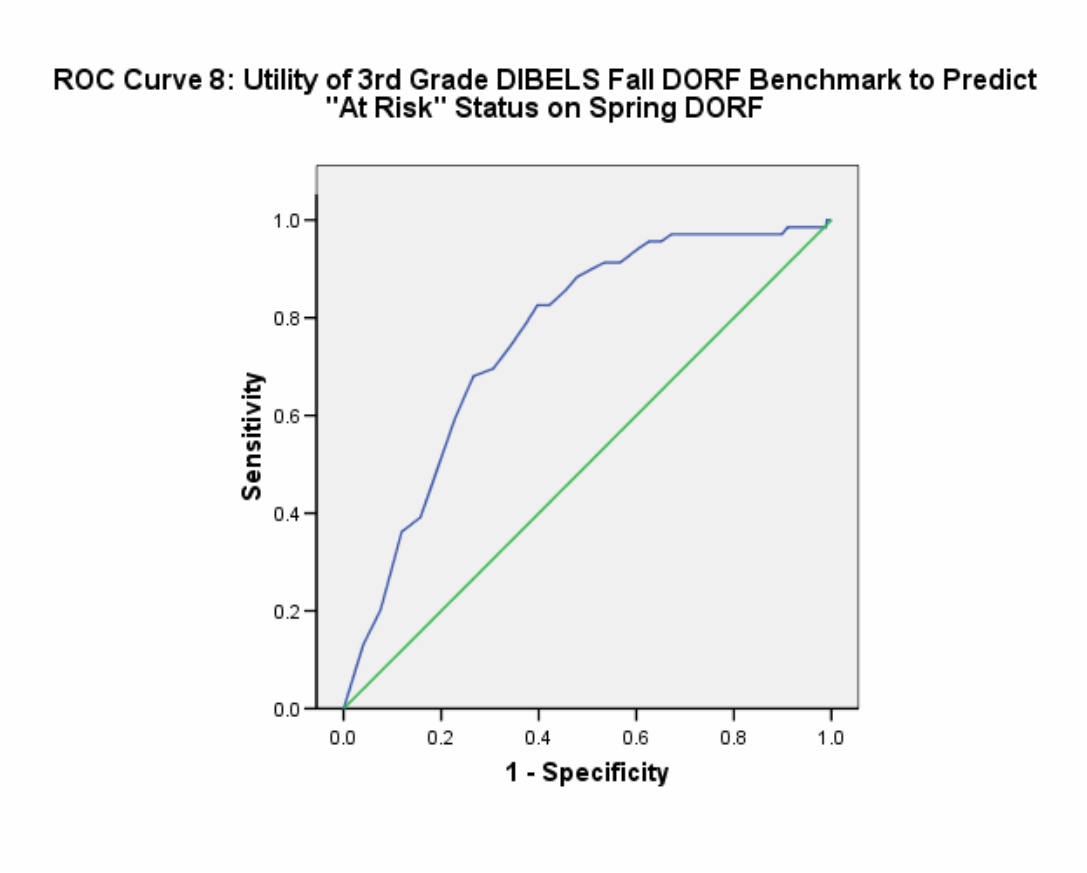

Figure 10: ROC Curve 8

Table 39: ROC Curve 8 Summary Table: 3rd Grade DIBELS

	DIBELS Cut-point	True Positive Decisions	False Positive Decisions
	53 cwpm	79%	20%
Area Under the Curve	**Recommended Cut-Point to achieve .75**	**Resulting True Positive Decisions**	**Resulting False Positive Decisions**
.82	NA	NA	NA

* Comparison of Sensitivity/Specificity Balance of the Author Recommended Cut-point for At Risk Status on the Fall DORF Subtest to the Sensitivity/Specificity Balance of a Newly Recommended Cut-Point for At Risk Status on the Fall DORF Subtest

The ROC curve for a Fall DORF cut point of 53 cwpm shows that it was an accurate predictor of "at risk" performance in the Spring overall. Although 538 of students were found not truly "at risk", the diagnostic accuracy and decision making utility estimates fall within acceptable parameters. While the area under the curve (area=.82) is not as large as the "low risk" predictive utility, it exceeds .75 and therefore indicates that the measure has an acceptable balance of sensitivity and specificity for identifying at-risk readers.

4.1.4 Research Questions 1d and 1e Results

Research Questions 1d and 1e

Are the established, author-recommended DIBELS cut-points for each benchmark period in grades one to three accurate predictors of achieving subsequent benchmark goals?

d) Did third graders who achieved DORF benchmark goals in the Spring achieve "Proficient or Advanced" on the PSSA?

e) Will the use of ROC analysis result in the establishment of cut-points with greater predictive power including an appropriate balance of sensitivity and specificity?

4.1.4.1 Descriptive Statistics

The descriptive characteristics of third grade 2006 DIBELS and PSSA performance data are presented in Table 40. Only records containing Spring DIBELS and PSSA results were analyzed.

Table 40: Descriptive Statistics: 3rd Grade Spring DIBELS and PSSA 2006

Measure	N	Mean DORF	DORF Range	DORF Std. Deviation	Mean Scaled Score PSSA	Scaled Score Range PSSA	PSSA Std. Deviation
Spring DORF & PSSA Reading	9215	96.85	0-253	36.46	1196	480-1999	229.53

Results of the evaluation of assumptions indicated no violations of homoscedacity or multicolinearity.

4.1.4.2 Correlation

Table 41: Correlation between 3rd Grade Spring DIBELS DORF and PSSA Reading Subtests

Measure	Spring DORF
Spring DORF	–
PSSA Reading	.71

The correlation between this DIBELS subtest and the Reading subtest of the state assessment is significant at the $p<.01$ level. Results indicate a strong relationship between students' Spring performance on the DIBELS DORF and their reading achievement on the PSSA.

4.1.4.3 Partial Replication Good et al. (2001) Study

Table 42 displays the performance consistency for 3rd grade students on the Spring DORF and Reading subtest of the PSSA in 2006. To parallel the 2001 Oregon study, only the proportion of students in the "low risk" and "at risk" categories on the DIBELS were calculated, "some risk" statistics were not considered. PSSA performance levels were explored in relationship to meeting or not meeting proficiency standards in that Proficient and Advanced categories were examined collectively as were Basic and Below Basic categories. In the table below column 1 lists the measures that were compared. Columns 2-5 report the percent of students: a) who achieved DIBELS benchmarks in the Spring who also achieved or exceeded

PSSA proficiency standards in the Spring) who did not achieve DIBELS benchmarks in the Spring and who did not meet PSSA proficiency levels; c) who achieved Spring DIBELS benchmarks but not PSSA proficiency; or d) who did not achieve Spring DIBELS benchmarks but experienced success on the PSSA.

Table 42: Consistency of 3rd Grade Students' DIBELS/PSSA Achievement in 2006

1	2	3	4	5
Measures	**% Students Low Risk DIBELS/ Proficient or Advanced PSSA**	**% Students At Risk DIBELS/ Not Proficient PSSA**	**% Students Low Risk DIBELS/ Not Proficient PSSA**	**% Students At Risk DIBELS/ Proficient or Advanced PSSA**
Spring DORF/ Spring PSSA	77%	72%	6%	10%

Column 2: Low Risk DIBELS Who Were Proficient or Above on PSSA

Three thousand three hundred thirty-six students reached or exceeded DIBELS DORF goals for "low risk" status in the Spring of 2006. Seventy-seven percent of those third graders also achieved proficiency on the PSSA that year (n=2,569). Consistent classifications occurred for struggling readers as well.

Column 3: At Risk DIBELS Who Were Not Proficient or Above on PSSA

Seventy-two percent of the 2,269 students reading below 52 words per minute (i.e. students "at risk") on the Spring DORF did not meet Reading proficiency standards on the PSSA.

Column 4: Low Risk DIBELS Who Were Not Proficient or Above on PSSA

Some students achieved surprising results on the PSSA however that raise questions about the accuracy of their DIBELS scores. Column 4 displays the results for the first group of interest. Sic percent of the students who demonstrated strong reading skills on the DORF (i.e. "low risk") struggled with the reading tasks on the PSSA (n=200).

Column 5: At Risk DIBELS Who Were Proficient or Above on PSSA

Moreover, 10% of the students who were flagged as "at risk" readers on the DORF achieved or exceeded proficiency on the PSSA. Contextually, 227 who were predicted likely to experience reading failure without intensive instructional support exceeded third grade performance standards in reading. The diagnostic accuracy and ROC statistics discussed below shed some light on these discrepancies.

4.1.4.4 Diagnostic Accuracy Analysis

The following 2x2 matrix illustrates the population of students discussed in the following results report. The comparison format is identical to Figure 2, however the prediction and outcome measures changed based on DIBELS to PSSA comparisons.

		Outcome Measure **3rd Grade PSSA Reading**	
Prediction Measure		Proficient or Advanced	Basic or Below Basic
3rd Grade Spring DIBELS DORF	Low Risk	*E*	*F*
	At Risk	*H*	*G*

E: students who achieved "low risk" status on DIBELS in the Spring who also achieved "proficient" or "advanced" status on the PSSA.

F: students who achieved "low risk" status on DIBELS in the Spring who achieved "basic" or "below basic" status on the PSSA.

H: students who achieved "at risk" status on DIBELS in the Spring who achieved "proficient" or "advanced" status on the PSSA.

G: students who achieved "at risk" status on DIBELS in the Spring who achieved "proficient" or "advanced" status on the PSSA.

Figure 11: Decision Matrix for DIBELS to PSSA Predictions

The percentages of students correctly or incorrectly identified as "low risk" in the Spring based on Fall DORF performance are listed in the table below.

Table 43: Diagnostic Accuracy of Third Grade DIBELS Spring DORF Benchmark to Predict "Proficient or Advanced" Performance on PSSA Reading

DORF Cut-point	**True Negative Decisions** ***(E)***	**False Negative Decisions** ***(F)***
110 cwpm	68%	15%

Results suggest that using the author recommended cut-point of 110 correct words read per minute, the Spring DIBELS oral reading fluency measure accurately classified 68% (n=2,268) of students as "low risk" in relationship to Proficient or Advanced performance on the PSSA. Those students fell into box *E* in Figure 11. However, 500 students were inaccurately labeled by DIBELS. For those students, DIBELS was not an accurate predictor of PSSA performance (box *F* Figure 11). Realistically, they may have benefited from additional reading support but did not receive it because of DIBELS inaccurate group placement.

Table 44: Diagnostic Accuracy of Third Grade DIBELS Spring DORF Benchmark

Cut-point	**True Positive Decision** ***(G)***	**False Positive Decision** ***(H)***
80 cwpm	63%	21%

The analysis indicated that more than half (63%) of the "at risk" students were accurately identified using a cut-point of 80 words correct per minute on the Spring DORF subtest when predicting Spring achievement on the PSSA. Therefore, 1,429 third graders in need of intensive reading instruction were truly "at risk" (box *G* Figure 11). Inaccurate identification occurred for 477 third graders. Those students fell into box H in Figure 11. Because the DIBELS classification was not an accurate predictor of Spring PSSA achievement, they were not appropriate candidates for receiving intensive reading support.

4.1.4.5 ROC Curve Analysis

While any number of students falsely identified seems educationally inappropriate, further analysis of two different ROC curves generated for the Spring DORF cut-points helped to determine the statistically appropriate balance of true positives vs. false positive identification of students in third grade. Decision criterion levels for selecting accurate cut-points were based on Swets (1988) recommendation that the area under the curve should be .75< to achieve an appropriate proportion of students.

Table 45: ROC Curve Descriptions for 3rd Grade Spring DIBELS to PSSA Comparisons

Curve	Outcome	Predictor Measure	Outcome Measure
9	Successful (Proficient or Advanced PSSA)	Spring DORF	PSSA Reading
10	Unsuccessful (Basic or Below Basic PSSA)	Spring DORF	PSSA Reading

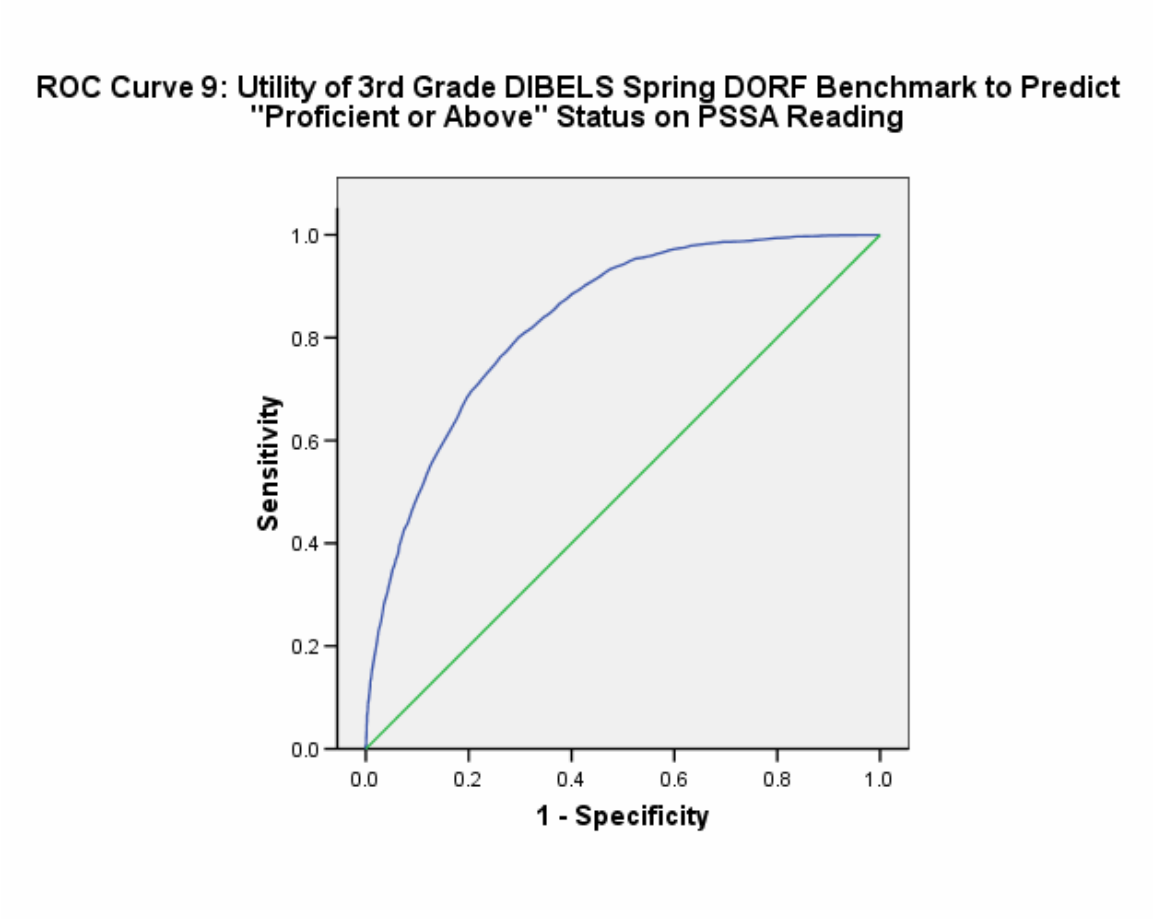

Figure 12: ROC Curve 9

Table 46: ROC Curve 9 Summary Table: 3rd Grade Spring DIBELS to PSSA 2006

	DIBELS Cut-point	True Positive Decisions	False Positive Decisions
	110 cwpm	68%	15%
Area Under the Curve	**Recommended Cut-Point to achieve .75**	**Resulting True Positive Decisions**	**Resulting False Positive Decisions**
.83	NA	NA	NA

* Comparison of Sensitivity/Specificity Balance of the Author Recommended Cut-point for Low Risk Status on the Spring DORF Subtest to the Sensitivity/Specificity Balance of a Newly Recommended Cut-Point for Low Risk Status on the Spring DORF Subtest

When using .75 as the referent, it is clear that the ROC curve produced using the currently established Spring third grade DORF cut point of 110 correct words per minute exceeded acceptability requirements (area=.83). This suggests that there was an appropriate balance of sensitivity and specificity in the classification rates using Spring DIBELS results to predict PSSA results. There was no need to establish new cut-points.

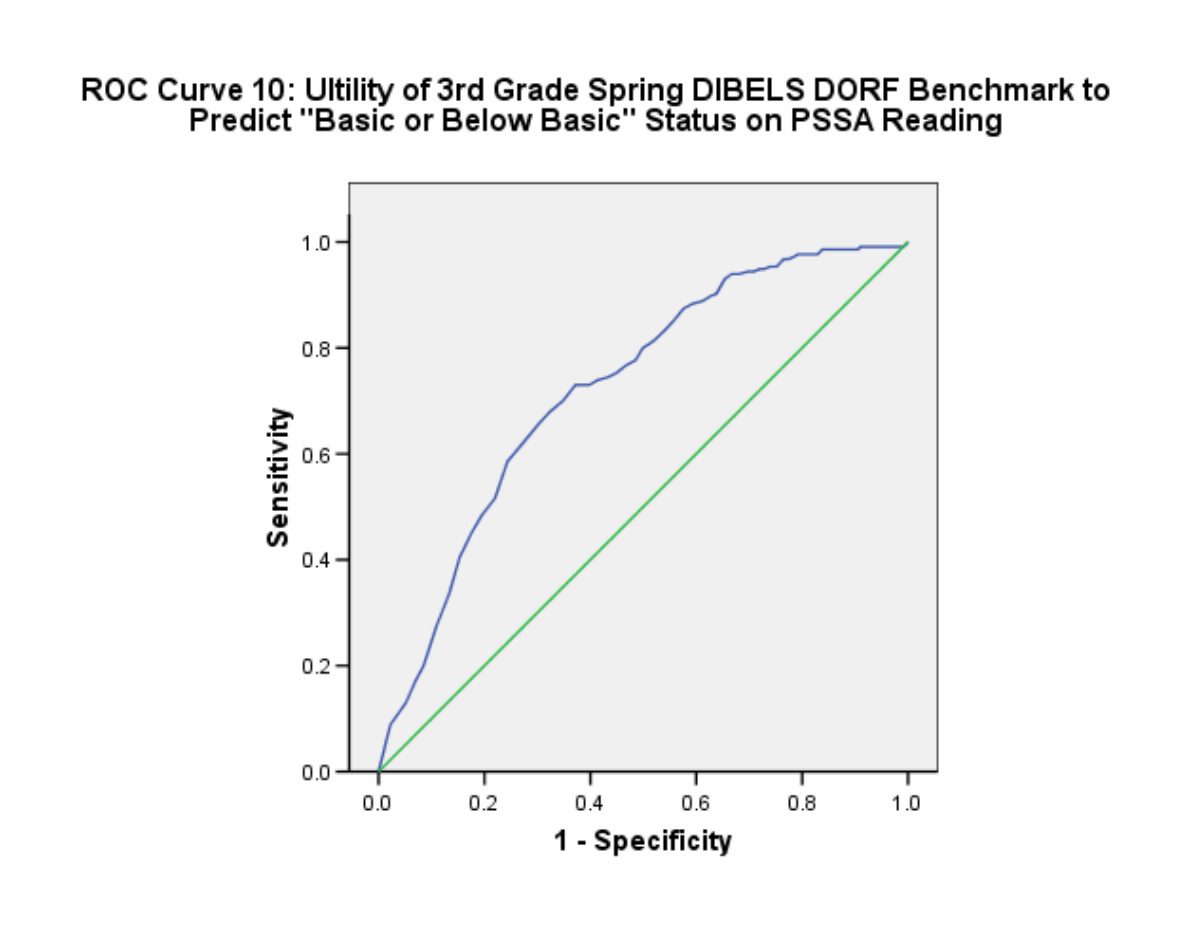

Figure 13: ROC Curve 10

Table 47: ROC Curve 10 Summary Table: 3rd Grade Spring DIBELS to PSSA 2006

	DIBELS Cut-point	True Positive Decisions	False Positive Decisions
	80 cwpm	63%	21%
Area Under the Curve	**Recommended Cut-Point to achieve .75**	**Resulting True Positive Decisions**	**Resulting False Positive Decisions**
.73	78 cwpm	67%	24%

* Comparison of Sensitivity/Specificity Balance of the Author Recommended Cut-point for At Risk Status on the Spring DORF Subtest to the Sensitivity/Specificity Balance of a Newly Recommended Cut-Point for At Risk Status on the Spring DORF Subtest

The ROC curve produced using the currently established cut point of 80 correct words read in one minute for students Spring performance on the DORF subtest did not meet acceptability requirements (area=.73) indicating an inappropriate balance of sensitivity and specificity. Analysis results suggest that to increase the area under the curve to a standard of .75, the cut point should be decreased by 2 words; thereby, increasing the accuracy of prediction to 67%. However, by making the measure more sensitive, it becomes less specific. The result would be an increase of false positive decisions by 3%. That means that 67 more students would have been inaccurately classified in the 2006 third grade sample. Those students would have been classified as "at risk" by DIBELS but exceeded performance standards on the PA state Reading test.

4.1.5 Research Questions 2a and 2c Results

Research Questions 2a and 2c

How accurately do the DIBELS measures administered in the Fall and Winter of first grade predict third grade Spring achievement on the DORF?

a) *What is the relationship between 3rd grade students' benchmark achievement on the 2006 Spring DORF DIBELS and their earlier first grade (2003/2004) DIBELS achievement on the Fall PSF, Fall NWF and Winter DORF subtests?*

c) *How much of the variance can be explained by the Level 2 variables built into the HLM design?*

   i) *DIBELS 1st Grade Fall PSF, Fall NWF, and Winter DORF scores*

   ii) *School*

   iii) *Student minority status*

   iv) *Student SES*

4.1.6 First Grade to Third Grade DIBELS Achievement Patterns

Table 48 reports the proportion of students achieving various benchmark levels on the subtests administered in September and January their first grade year and the DORF subtest administered in May of their third grade year. Only the proportion of students in the "low risk" and "at risk" categories were calculated, "some risk" statistics were not considered because the likelihood of either successful or unsuccessful future achievement is uncertain. Column 1 indicates that the students' first grade status on the PSF, NWF and DORF subtests in 2003/2004 are being compared to their third grade status on the DIBELS DORF in 2006. Columns 2-5 include the percent of students: a) who achieved DIBELS benchmarks in the first grade and in third grade; b) who did not achieve DIBELS benchmarks in first grade or third grade; c) who achieved first grade DIBELS benchmarks but not third grade DIBELS benchmarks; or d) who did not achieve first grade benchmarks but later achieved third grade benchmarks.

Table 48: Comparison of Students' 1st to 3rd Grade DIBELS Achievement from 2004 to 2006

1	2	3	4	5
Measures	**% Students Low Risk DIBELS 1st/ Low Risk DIBELS 3rd**	**% Students At Risk DIBELS 1st/ At Risk DIBELS 3rd**	**% Students Low Risk DIBELS 1st/ At Risk DIBELS 3rd**	**% Students At Risk DIBELS/ Low Risk DIBELS 3rd**
Fall 1st Grade PSF/ Spring 3rd Grade DORF	24%	21%	13%	22%
Fall 1st Grade NWF/ Spring 3rd Grade DORF	28%	24%	9%	20%
Winter 1st Grade DORF/ Spring 3rd Grade DORF	33%	24%	10%	19%

Column 2: Low Risk Who Stayed Low Risk

Twenty-four percent of the 3,816 first grade students who scored in the "low risk" range on the Fall PSF subtest experienced continued success on the DIBELS at the end of third grade evidenced by their "low risk" status on the Spring DORF (n=920). Similarly, 28% of the students who were "low risk" early in first grade on the NWF were "low risk" by the end of third grade on the DORF (1,159 of 4,117 students). Winter oral reading fluency status in first grade was consistent with third grade Spring comparisons for 33% of students. Those 1,529 students who read at least 20 words per minute on first grade text on the Winter DORF in first grade met or

exceeded proficiency standards for fluency (110 wpm on third grade text) by Spring of third grade.

Column 3: At Risk Who Stayed At Risk

Twenty-one percent of the students who did not easily segment phonemes at the end of first grade (i.e. students who were "at risk") did not read with adequate levels of fluency two years later (n=621 out of 2,896 students). NWF to DORF comparisons indicated that 760 (24%) of the 3,172 students who performed well below benchmark in the Fall of first grade performed well below benchmark by the Spring of third grade. Likewise, 24% of the first grade students who were "at risk" on the DORF assessment in Winter of 2004 were still "at risk" on the DORF assessment in Spring of 2006 (n= 560 of 2,362 students).

Column 4: Low Risk Who Became At Risk

There was significant movement between risk categories across the three years of schooling. Thirteen percent of students initially considered "low risk" by DIBELS PSF standards in first grade scored "at risk" on the Spring DORF subtest in third grade. That means that 505 of the 3,816 first grade students who were believed highly likely to achieve future DIBELS benchmark goals did not. The same discrepancies occurred on the NWF subtest. Three hundred seventy-two (9%) students who surpassed performance standards on the NWF in first grade failed to meet third grade DORF goals. Similarly, when comparing fluency growth, 10% of the 4,594 students labeled "low risk" on the Winter DORF subtest in first grade became "at risk" by the end of third grade on the DORF (n=437).

Column 5: At Risk Who Became Low Risk

Uneven achievement patterns also occurred for first grade students who were initially "at risk" on the DIBELS subtests administered early in the 2003-2004 school year. Twenty-two percent of the 2,986 students who scored well below benchmark on the Fall first grade PSF measures scored above benchmark on the Spring third grade DORF measure two years later (n=651). Slightly fewer (20%) of the 3,172 students who were "at risk" on the NWF subtest in first grade became "low risk" on the DORF subtest in third grade. Finally, 19% of the students labeled "at risk" for future reading difficulties due to poor performance on the DORF in January of their first grade year were "low risk" on the DORF by Spring of third grade (n= 442).

4.1.6.1 Correlation

A Pearson product moment correlation matrix was calculated to determine the strength of the relationships between students' first grade DIBELS performance and their later performance on the DORF in third grade. Coefficients were also calculated to determine the strength of the linkage between students' school, minority status, and socioeconomic status and outcomes on the third grade fluency measure. The resulting correlation coefficients also informed the entry order for the HLM model.

Table 49: Correlation Across 1st to 3rd Grade DIBELS DORF with Level 2 Variables

Measure	Spring 3rd Grade DORF	Fall 1st Grade PSF	Fall 1st Grade NWF	Winter 1st Grade DORF	School	Minority Status
Spring 3rd Grade DORF	_					
Fall 1st Grade PSF	.17	_				
Fall 1st Grade NWF	.26	.59	_			
Winter 1st Grade DORF	.38	.52	.67	_		
School	.01	.12	.11	.13	_	
Minority Status	-.19	-.15	-.10	-.10	-.10	_
SES	-.18	-.19	-.19	-.10	-.19	.45

All results are significant at the $p<.01$ level. When examining early predictors of reading achievement, we see that first grader's scores on the NWF subtest in the Fall were more strongly linked to their eventual achievement on the Spring DORF in third grade than were PSF scores (.26 & .17 respectively). Overall, students' performance on the DORF subtest in the Winter of first grade correlated the strongest with later DORF performance in the Spring of third grade (.38). Minority status produced a correlation of -.19 which was the strongest of the three variables related to group membership. Only slightly less was students' socioeconomic status (-.18). The negative correlations reported for these variables indicate that students who are not

economically disadvantaged and students who are not minorities performed better on the DIBELS than students who are economically disadvantaged or racial minorities.

4.1.6.2 HLM Growth Analysis

HLM provides an accurate measure of change and growth over time because it facilitates multiple-time-point analysis and accounts for individual variability as well as under/over estimation of observed relationships (Raudenbush & Bryk, 2002). This model was particularly informative when examining the accuracy of DIBELS predictions for long-term literacy achievement for students in Reading First schools. The HLM procedure was particularly useful in analyzing this longitudinal data set because it accommodated for missing data by estimating missing data points from existing data to conduct the analysis. The model below outlines the entry order of the predictor variables for the HLM analysis. More highly correlated variables were entered first to test and correct for significant effects due to colinearity.

Table 50: Growth Model Analysis 1

Model #	Outcome Variable	Predictor Variable
1	Spring 3rd DIBELS DORF	• Winter 1st Grade DIBELS DORF • Fall 1st Grade DIBELS NWF • Fall 1st Grade DIBELS PSF • Student Minority Status • Student SES • School

Growth estimates were based on individual student parameters computed by the HLM program. Individual growth estimates were calculated for students' achievement on the Fall and Winter 1st Grade DIBELS subtests (PSF, NWF, DORF) compared to later achievement on the 3rd Grade

DORF. The analysis confirmed the moderate correlation between first grade Winter DORF scores and 3rd grade Spring DORF scores (R=.38). The R^2 value indicated that only 15% of the variance in students' 3rd grade scores was accounted for by their earlier 1st grade scores. To determine the extent to which the Level 2 variables contribute to this variance, each was entered into the model in the order noted above. The results of the HLM analysis including all Level 2 predictors appear in the table below.

Table 51: HLM Results Examining the Combined Influence of all Predictor Variables for 1st Grade to 3rd Grade DIBELS Achievement

Variable	R	R^2	EB	p value
Winter 1st Grade DIBELS DORF	.38	.15	.39	0.000
Fall 1st Grade DIBELS NWF	.41	.17	.15	0.000
Fall 1st Grade DIBELS PSF	.42	.18	.08	0.000
Student SES	.45	.20	.13	0.000
Student Minority Status	.48	.23	.15	0.000
School	.50	.25	.13	0.000

The analysis showed that all Level 2 variables had a statistically significant impact on end-of-third grade DIBELS performance at the $p<.01$ level. As the additional Level 2 predictor variables were added into the model, the strength of the correlation progressively increases. For example, both first grade Winter DORF scores and Fall NWF produced a coefficient of .41suggesting that these variables combined were more predictive of end-of-third grade

outcomes than students DORF scores alone. Results on the PSF subtest only minimally contributed. The strength of the predictive relationship for first to third grade achievement outcomes on the DIBELS was also impacted by students' minority status, SES, and school. The correlation steadily increased as each variable was entered into the model. Overall, the combined influence of all Level 2 variables accounted for 25% of the variance in students' 3rd grade DORF scores.

The EB values (column 4 Table 51) represent the *Empirical Bayes Estimates* produced in the analysis, which indicate the individual influence of each of the entered predictors in the HLM model. Overall, first grade Winter DORF scores exerted the largest influence (.39). Moreover, the influence of the other Level 2 variables was relatively constant with each being related minimally to third grade achievement. With all Level 2 variables entered into the model, the weakest relationship occurred for first grade PSF subtest scores with a coefficient of only .08.

4.1.7 Research Questions 2b and 2c Results

Research Questions 2b and 2c

How accurately do the DIBELS measures administered in the Fall and Winter of first grade predict third grade Spring achievement on the DORF and on the PSSA?

b) What is the relationship between 3rd grade students' achievement on the 2006 PSSA Reading subtest and their earlier first grade (2003/2004) DIBELS achievement on the Fall PSF, Fall NWF and Winter DORF subtests?

c) How much of the variance can be explained by the Level 2 variables built into the HLM design?

i) DIBELS 1st Grade Fall PSF, Fall NWF, and Winter DORF scores

ii) School

iii) Student minority status

iv) Student SES

4.1.7.1 First Grade DIBELS to Third Grade PSSA Achievement Patterns

Table 52 summarizes the consistency of students performance on the DIBELS PSF, NWF, and DORF subtests in the Fall/Winter of first grade (2003/2004) compared to their performance on the reading subtest of the PSSA in the Spring of third grade (2006). In the table below Column 1 indicates that students' performance on various DIBELS subtests in first grade was compared to their performance on the PSSA in third grade. Columns 2-5 report the percent of students: a) who achieved DIBELS benchmarks in first grade who also achieved or exceeded PSSA proficiency standards in the Spring of third grade b) who did not achieve DIBELS benchmarks in first grade and who did not meet PSSA proficiency levels in the Spring of third grade; c) who achieved DIBELS benchmarks in first grade but not PSSA proficiency in the Spring of third grade; or d) who did not achieve DIBELS in first grade benchmarks but experienced success on the PSSA in third grade.

Table 52: Comparison of Students' 1st to 3rd Grade DIBELS to PSSA Achievement from 2004 to 2006

1	2	3	4	5
Measures	**% Students Low Risk DIBELS/ Proficient or Advanced PSSA**	**% Students At Risk DIBELS/ Not Proficient PSSA**	**% Students Low Risk DIBELS/ Not Proficient PSSA**	**% Students At Risk DIBELS/ Proficient or Advanced PSSA**
Fall 1st Grade PSF/ 3rd Grade PSSA	18%	42%	19%	19%
Fall 1st Grade NWF/ 3rd Grade PSSA	20%	44%	20%	19%
Winter 1st Grade DORF/ 3rd Grade PSSA	23%	40%	20%	13%

Column 2: Low Risk DIBELS Who Were Proficient or Above PSSA

Three thousand eight hundred sixteen first graders met or exceeded benchmark goals on the PSF test in the Fall of 2003. Of those students, 18% went on to achieve "proficient or advanced" status on the 3rd grade PSSA reading test in the Spring of 2006 (n=683). Eight hundred twelve (20%) of the students who were "low risk" on the Fall NWF measure in first grade also performed well on the PSSA by the end of third grade. That consistent pattern of successful reading achievement also occurred for 23% of the 4,594 students who exceeded oral reading fluency goals in January of first grade. Specifically, 1,069 of the students who were

labeled “low risk” by DIBELS performance standards on the Winter DORF in first grade were also “proficient or advanced” on the PSSA 2.5 years later.

Column 3: At Risk DIBELS Who Were Not Proficient or Above PSSA

Long-term patterns of achievement were evident for struggling readers as well. Forty-two percent of the 2,986 students who scored “at risk” on the PSF at the beginning of first grade performed below proficiency standards on the PSSA in third grade (n=1,251). Consistent low achievement occurred for students “at risk” on the NWF subtest as well. Forty-four percent of the 3,172 students who scored significantly below benchmark standards on the DIBELS in 2003 scored at “basic” or “below basic” levels on the PSSA in 2006 (n= 1,407). Nine hundred forty eight (48%) students followed suit for DIBELS DORF to PSSA comparisons from 1st to 3rd grade.

Column 4: Low Risk DIBELS Who Were Not Proficient or Above PSSA

Considerable movement occurred between reading achievement categories across the three years studied. Nineteen percent of the students who scored at “low risk” levels on the PSF in first grade struggled on the reading subtest of the PSSA in third grade (n=717). The same drop in status occurred for 20% of the students initially considered “low risk” on the NWF and on the DORF (n=816 & 913 respectively). DIBELS was not an accurate early predictor of future reading achievement for those students. Fall/Winter 2003/2004 DIBELS results suggested that these students should have progressed successfully with reading over time but they did not.

Column 5: At Risk DIBELS Who Were Proficient or Above PSSA

Unlike the sample of students discussed earlier, surprising percentages of students flagged for reading failure by DIBELS first grade subtests exceeded performance goals on the PSSA by Spring of third grade. Nineteen percent of the 2,896 students scoring "at risk" on the PSF subtest in the Fall of 2003 achieved proficiency on the PSSA in 2006. Moreover, 595 (19%) students who were unlikely to achieve subsequent benchmark goals (i.e. students "at risk" for reading difficulties) based on low NWF scores in first grade exceeded PSSA performance standards in third grade. Finally, DORF to PSSA comparisons showed that 13% of the "at risk" group achieved "proficient" or "advanced" status by the end of third grade.

4.1.7.2 Correlation

Pearson product moment correlations were calculated to determine the strength of the relationships between students' performance on selected DIBELS subtests in first grade and their later performance on the Reading PSSA in third grade. Coefficients were also calculated to determine the strength of the linkage between students' school, minority status, and socioeconomic status and outcomes on the third grade PSSA. The resulting correlation coefficients informed the entry order of the HLM model.

Table 53: Correlation Across 1st Grade DIBELS DORF to 3rd Grade PSSA Reading with Level 2 Variables

Measure	3rd Grade PSSA Reading	Winter 1st Grade DORF	Fall 1st Grade NWF	Fall 1st Grade PSF	School	Minority Status
3rd Grade PSSA Reading	_					
Winter 1st Grade DORF	.40	_				
Fall 1st Grade NWF	.30	.59	_			
Fall 1st Grade PSF	.26	.52	.67	_		
School	-.03	.12	.11	.13	_	
Minority Status	-.24	-.15	-.10	-.10	-.10	_
SES	.30	-.19	-.19	-.10	-.19	.45

Results showed that there is a moderate relationship between first grade DIBELS DORF performance and 3rd grade PSSA performance (.40). First grade Fall DIBELS NWF scores were stronger predictors of reading scores on the PSSA in third grade (r=.30) than were Fall PSF scores (r=.40). Interestingly, students' economic status was as strongly correlated to students' performance on the state test in third grade (r=.26) as was their performance on DIBELS PSF in the Fall of first grade (-.26). Similar to the coefficients for DIBELS outcomes, the negative correlations reported for students' minority status and SES indicate that students who are not minorities and students who are not economically disadvantaged perform better on the PSSA

than students who are minorities and students who are economically disadvantaged. School had the weakest relationship to students' 3rd grade PSSA scores (.03).

4.1.7.3 HLM Growth Analysis

The model below outlines the entry order of the predictor variables for the HLM analysis.

Table 54: Growth Model Analysis 2

Model #	Outcome Variable	Predictor Variable
2	Spring 3rd PSSA Reading	• Winter 1st Grade DIBELS DORF • Fall 1st Grade DIBELS NWF • Fall 1st Grade DIBELS PSF • Student SES • Student Minority Status • School

Individual growth estimates were calculated for students' achievement on the Fall 1st Grade DIBELS PSF and NWF subtests as well as the 1st Grade Winter DORF subtest then compared to 2006 achievement on the 3rd grade Reading subtest of the PSSA. Table 56 shows the moderate correlation between first grade DORF scores and 3rd grade PSSA scores (R=.40). Sixteen percent of the variance in students' PSSA achievement was explained by their first grade oral reading proficiency in January of 2004. To determine the extent to which the Level 2 variables contribute to this variance, each was entered into the model in the order noted above. The results of the HLM analysis including all Level 2 predictors appear in the table below. "School" did not meet entry requirements into the model. It was not a significant predictor in end-of-third grade DORF outcomes.

Table 55: HLM Results Examining the Combined Influence of all Predictor Variables for 1st Grade DIBELS to 3rd Grade PSSA Achievement

Variable	R	R^2	EB	p value
Winter 1st Grade DIBELS DORF	.40	.16	.39	0.000
Fall 1st Grade DIBELS NWF	.42	.18	.13	0.000
Fall 1st Grade DIBELS PSF	.42	.18	.03	0.090
Student SES	.44	.19	.13	0.000
Student Minority Status	.45	.20	.09	0.000

As the additional predictor variables were added into the model, the strength of the correlation gradually increased. A moderate correlation existed between all four Level 2 variables and students' achievement on the PSSA (.45). The combined influence of all remaining Level 2 variables accounts for only 20% of the variance in students' 3rd grade PSSA scores. PSF did not add to the correlation nor increase the amount of variance explained.

The Empirical Bayes Estimate (EB) values produced in the analysis indicate the individual influence of each of the entered predictors with all Level 2 variables in the model. Overall, first grade Winter DORF scores exerted the largest influence (EB=.39). At the $p<.01$ significance level, the PSF subtest scores did not impact 3rd grade reading outcomes (EB=.03; p=.09). Interestingly, students' socioeconomic status impacted students PSSA achievement equally as much as their first grade performance on the NWF subtest did (EB=.13).

4.1.8 Research Question 3a and 3b Results

Research Question 3a and 3b

Are the DIBELS subtests administered in the Fall and Winter of first grade accurate predictors of eligibility for special education services in reading at the end of third grade?

a) *What is the relationship between students' eligibility status at the end of third grade and earlier first grade DIBELS benchmark status of "at risk" on the Fall PSF & NWF subtests and Winter DORF subtest?*

b) *What is the relationship between students' eligibility status at the end of third grade and their minority and SES status?*

4.1.8.1 Consistency of DIBELS Classification and Special Education Status

Table 56 reports the proportion of students achieving "low risk" or "at risk" status on DIBELS subtests early in first grade and their special education status at the end of third grade. To focus on 3rd grade eligibility status that related closely to reading difficulties, only high incidence primary disability classifications were coded as 1 (IEP). The following disability categories were not coded as IEP: speech/language impairment (n=112), other health impairments (n=21), visual impairment (n=7), multiple disabilities (n=6), orthopedic impairments (n=4), autism (n=4), and hearing impairment (n=3).

In the Table 56, column 1 indicates that students' first grade performance on the DIBELS PSF, NWF and DORF subtests in the Fall/Winter of 2003/2004 were being compared to their third grade special education status (IEP/No IEP). Columns 2-5 include the percent of students: a) who achieved DIBELS benchmarks in the first grade and were not eligible for special education in third grade; b) who did not achieve DIBELS benchmarks in first grade and were

eligible for special education in third grade; c) who achieved first grade DIBELS benchmarks but were eligible for special education in third grade; or d) who did not achieve first grade benchmarks but were not eligible for special education in third grade.

Table 56: Comparison of Students' 1st Grade DIBELS At Risk Status and 3rd Grade Special Education Status from 2004 to 2006

1	2	3	4	5
Comparisons	**% Students Low Risk DIBELS 1st/ No IEP 3rd**	**% Students At Risk DIBELS 1st/ IEP 3rd**	**% Students Low Risk DIBELS 1st/ IEP 3rd**	**% Students At Risk DIBELS/ No IEP 3rd**
Fall 1st Grade PSF/ Spring 3rd Grade IEP	44%	12%	8%	55%
Fall 1st Grade NWF/ Spring 3rd Grade IEP	48%	13%	5%	32%
Winter 1st Grade DORF/ Spring 3rd Grade IEP	53%	15%	6%	27%

Column 2: Low Risk DIBELS Who Were Not Eligible for Special Education

According to DIBELS authors, established subtest performance ranges indicate the likelihood of students' either experiencing future reading success or failure. 2001 study results suggest that it is a valid assessment for accurate identification of students' reading difficulties with a primary purpose of determining children at-risk for reading failure (Good, Kaminski, Simmons, & Kame'enui). Therefore, if predictions are accurate, students scoring in the "low risk" range in first grade will continue to experience reading success in later school years.

The longitudinal data analyzed for this study indicate that 44% of the students who scored in the "low risk" range on the PSF subtest in the Fall of first grade were not eligible for special education services in reading by Spring of third grade (n=1,697 out of 3,816 students). Likewise, DIBELS prediction of "low risk" for reading difficulties held true for the majority of students who exceeded benchmark goals on the NWF in the Fall of first grade. Those 1,996 (48%) high performing students continued to succeed in the general education curriculum for reading in third grade. Finally, as expected 53% of the 4,594 students who were "low risk" on the DIBELS DORF in the middle of first grade were not eligible for special education by the end-of third grade.

Column 3: At Risk DIBELS Who Were Eligible for Special Education

Again, if DIBELS prediction of reading difficulties are accurate, students scoring in the "at risk" range in first grade would continue to experience reading failure in later school years (unless effective intensive intervention occurred), which would likely lead to identification for special education. In the sample studied, 12% of the 2,986 students flagged as "at risk" on the PSF subtest in the Fall of first grade were eligible for special education in third grade. Thirteen percent of the "at risk" students on the Fall NWF subtest were also identified by third grade (n= 421). According to oral reading fluency indicators, 15% of the 2,362 first grade students who were at risk in January of 2004 were receiving special education services by the end of third grade (n=345).

Column 4: Low Risk DIBELS Who Were Eligible for Special Education

Some students meeting or exceeding DIBELS goals on the PSF, NWF, and DORF subtests administered early in first grade were ultimately identified for special education in third

grade. Specifically, 319 (8%) of the students who were "low risk" on the PSF in the Fall of first grade had IEPs in third grade. Two hundred twenty-one (5%) students initially considered "low risk" by DIBELS NWF standards were eligible for special education support in third grade. Likewise, 6% of the students who scored "low risk" on the DIBELS DORF in first grade were listed as having IEPs in third grade (n=254). Those patterns of under-identification suggest that DIBELS results were not accurate indicators of future reading achievement for those students.

Column 5: At Risk DIBELS Who Were Not Eligible for Special Education

The proportion of students predicted to be struggling readers (i.e. students "at risk") by first grade DIBELS performance categories who did not end-up in special education in third grade suggest high over-identification rates. That is, considerable numbers of students flagged as "at risk" did not experience enough difficulty in the general education curriculum to be eligible for special education services in third grade. One thousand six hundred twenty-nine (55%) of the students who scored in the "at risk" range on the PSF subtest at the beginning of first grade were not identified for special education by third grade. Thirty-two percent of students considered "at risk" on the NWF subtest in first grade were not eligible for special education services two years later (n=1,015). Moreover, DORF results showed that 27% of the 2,362 "at risk" first graders did not need special education support in third grade (n=638).

4.1.8.2 Diagnostic Accuracy Analysis

The following 2x2 matrix illustrates the population of students discussed in the following results report. The comparison format is identical to Figures 2 & 11, however, the prediction measures included first grade DIBELS PSF, NWF and DORF subtests and the outcome of interest was third grade disability status.

		Outcome Measure **3rd Grade Special Education Status**	
		No IEP	IEP
Prediction Measure **1st Grade Fall DIBELS PSF**	Low Risk	*I*	*J*
	At Risk	*L*	*K*

I: students who achieved "low risk" status on DIBELS in the first grade who were not eligible for special education in third grade.

J: students who achieved "low risk" status on DIBELS in first grade who were eligible for special education in third grade.

L: students who achieved "at risk" status on DIBELS in first grade who were not eligible for special education in third grade.

K: students who achieved "at risk" status on DIBELS in first grade who were eligible for special education in third grade.

Figure 14: Decision Matrix for DIBELS to IEP Predictions

The percentages of students correctly or incorrectly identified for special education are listed below.

Table 57: Diagnostic Accuracy of Fist Grade DIBELS Fall PSF Benchmark to Predict Third Grade Status of No IEP

Cut-point	**True Negative Decisions** ***(I)***	**False Negative Decisions** ***(J)***
35 cspm	68%	13%

Results suggest that using the author recommended cut-point of 35 correct sounds produced per minute, the DIBELS PSF administered in the Fall of first grade accurately classified 68% of the

3,816 students as "No IEP" in third grade. Those students fell into box *I* in Figure 14 and became good readers as expected (n=2,594). However, 13% of first grade students were inaccurately labeled by the Fall DIBELS PSF subtest. For those 496 students, DIBELS was not an accurate predictor of third grade IEP status (box *J* Figure14). Those students were ultimately identified as having a disability. They may have benefited from additional reading support and early intervention to alter their achievement trajectories but would not have been identified as needing that support based on their first grade Fall PSF performance.

Table 58: Diagnostic Accuracy of Fist Grade DIBELS Fall PSF Benchmark to Predict Third Grade Status of IEP

Cut-point	**True Positive Decisions** ***(K)***	**False Positive Decisions** ***(L)***
10 cspm	22%	37%

Two thousand nine hundred eighty-six students were categorized as "at risk" on the Fall PSF subtest. Twenty-two (n=657) were accurately identified using a cut-point of 10 correct sounds per minute on the Fall PSF test to predict third grade Spring outcomes (box *K* in Figure 14). Inaccurate identification occurred for 37% of the participating first graders who were initially labeled "at risk." Ultimately 1,075 students were not eligible for special education (box *L* in Figure 14). In the prevention framework of Reading First, those students probably received additional reading support but may not have needed the intervention.

Table 59: Diagnostic Accuracy of First Grade DIBELS NWF Benchmark to Predict Third Grade Status of No IEP

Cut-point	True Negative Decisions *(I)*	False Negative Decisions *(J)*
24 cspm	70%	12%

Four thousand one hundred seventeen students achieved benchmark on the Fall NWF subtest in first grade. The results show that 70% (n=2,882) of those "low risk" students were accurately identified as 'good readers'(i.e. non-special education) using a cut-point of reading at least 24 correct sounds Conversely, 12% of "low risk" students were identified incorrectly; the DIBELS results were not accurate indicators of future successful achievement for 494 students. Instead, those students slotted for reading success ended up in special education by third grade.

Table 60: Diagnostic Accuracy of First Grade DIBELS NWF Benchmark to Predict Third Grade Status of IEP

Cut-point	True Positive Decisions *(K)*	False Positive Decisions *(L)*
13 cspm	40%	23%

The analysis indicated that 40% of the 3,172 "at risk" students were accurately identified using a cut-point of 13 correct sounds per minute on the Fall NWF test to predict Spring outcomes. Those students exhibited the consistent achievement patterns explained in box *K* of Figure 14. In stark contrast, 729 students exhibited unexpected achievement. Inaccurate identification occurred for 23% of those first graders. Although the students in box *L* were categorized "at risk" in first grade, they did not qualify for special education services in third grade.

Table 61: Diagnostic Accuracy of First Grade Winter DIBELS DORF Benchmark to Predict Third Grade Status of No IEP

Cut-point	True Negative Decisions *(I)*	False Negative Decisions *(J)*
20 wpm	70%	15%

Four thousand five hundred ninety-four students scored in the "low risk" range on the DORF using the performance criterion of 20 words read correctly per minute in January of first grade. Results suggest that 3,216 (70%) of those students were accurately classified as good readers because they continued to experience reading success later in school (box *I* Figure 14). However, 689 students were inaccurately identified by DIBELS. Those students were identified as having a disability likely requiring reading intervention by the end of third grade.

Table 62: Diagnostic Accuracy of First Grade Winter DIBELS DORF Benchmark to Predict Third Grade Status of IEP

Cut-point	True Positive Decisions *(K)*	False Positive Decisions *(L)*
8 cwpm	60%	39%

The analysis indicated that 60% of the 2,362 "at risk" students were accurately identified using a cut-point of 8 correct words per minute on the Winter DORF subtest in first grade to predict third grade disability status (n=1,417). Those students exhibited the consistent achievement patterns explained in box *K* of Figure 14. However, inaccurate identification occurred for 39% of that first grade sample. That means that 921 students labeled "at risk" for reading difficulties in first

grade did not develop disabilities by third grade. Their inconsistent achievement mirrors the patterns depicted in box *L* of Figure 14.

4.1.8.3 Correlation

Pearson product moment correlations were calculated to determine the strength of the relationships between students' performance on selected DIBELS subtests in first grade, their minority status, and their SES to their special education status in third grade.

Table 63: Correlation Across 1st Grade DIBELS DORF to 3rd Grade IEP Status

Measure	3rd Grade IEP Status	Fall 1st Grade PSF	Fall 1st Grade NW	Winter 1st Grade DORF	Minority Status
3rd Grade IEP Status	_				
Fall 1st Grade PSF	.13	_			
Fall 1st Grade NWF	.18	.59	_		
Winter 1st Grade DORF	.24	.52	.67	_	
Minority Status	.42	.-.15	-.10	-.10	_
SES	.34	-.19	-.19	-.10	.45

The data indicate that the relationship between first grade DIBELS achievement and special education eligibility in third grade is relatively weak with correlations ranging from .13 to .24. Of the three subtests examined, DIBELS DORF produced the highest correlation of .24. Students' minority status was correlated more strongly than all variables analyzed meaning that

minority students were more likely to be identified for special education than white students (.42). SES produced the second highest relationship (.34). Again, poor students were more often identified for special education than were students who were not economically disadvantaged.

4.1.8.4 Logistic Regression Analysis

Logistic regression was used to analyze the results given the dichotomous nature of the dependent variable (3rd grade special education status= IEP/No IEP). The results report the probability of students being identified for special education in third grade based on the first grade variables of DIBELS "at risk" status on the Fall PSF, Fall NWF, and Winter DORF subtests, minority status, and socio economic status.

Table 64: Logistic Results Predicting 3rd Grade Special Education Status Based on First Grade DIBELS Achievement, SES, and Minority Status

Variable	Exp (B)	p
Fall 1st Grade DIBELS PSF "at risk"	.34	.042
Fall 1st Grade DIBELS NWF "at risk"	1.2	.007
Winter 1st Grade DIBELS DORF "at risk"	1.7	.003
Student SES	2.1	.000
Student Minority Status	3.8	.000

The data show that with the exception of the DIBELS Fall PSF in first grade, all other variables were identified as unique predictors of special education eligibility in third grade ($p<.01$). Interestingly, students' socioeconomic status and minority status were more significant predictors than were students' achievement on the DIBELS. Probability estimates suggest that students' who were not white were 3.8 times more likely to be identified as having a disability in third grade than students who were white. Likewise, the odds of ending up in special education were 2.1 times greater for poor students than for students from economically secure homes. The odds ratios based on DIBELS performance were less significant. However, first grade "at risk" status on the Winter DIBELS DORF was more predictive of third grade eligibility than NWF achievement (Exp (B)=1.7).

5.0 CHAPTER FIVE

5.1 DISCUSSION

The purpose of this study was to examine the predictive strength and decision-making utility of the Dynamic Indicators of Early Literacy Skills (DIBELS; Good & Kaminski, 2002). Specifically, the study examined whether DIBELS benchmarks correctly differentiated among students who were at-risk for reading failure and those who were not as measured by end-of-grade achievement on the Pennsylvania System of School Assessment (PSSA) in Pennsylvania Reading First schools. More broadly, this study addressed the effectiveness of DIBELS for educational decision-making using the author-recommended benchmarks for early identification of children considered to be at-risk for reading failure. Additionally, 3rd grade special education eligibility data were analyzed to determine whether first grade Fall and Winter DIBELS cut points were appropriately sensitive and specific in relation to long-term predictions of special education outcomes in third grade. Taken together the results of these analyses contribute to the research base measuring the effectiveness of using DIBELS in an early intervention framework to accurately target students who are at risk for reading difficulties before their reading trajectories become intractable.

5.1.1 DIBELS Prediction of Short-term Achievement Outcomes

Results suggested that DIBELS performance indicators of "low risk" and "at risk" were generally predictive of first through third grade students' Fall to Spring achievement in 2006. That is to say that the majority of students who performed well on the DIBELS in the Fall performed well on the DIBELS in the Spring. The inverse was true for low-performing students. In general, older students were more consistent in their DIBELS performance than their younger peers. Higher percentages of students continued to score as either "low risk" or "at risk" on Fall to Winter to Spring test administrations than students who moved between achievement categories (e.g., "low risk" to "at risk"). A possible explanation for students' less erratic performance over time is that children's achievement patterns become more predictable and stable as children age (Fletcher et al. 2002). On the surface, DIBELS risk classifications were accurate predictors of future (albeit short-term) benchmark achievement for the majority of the students tested.

However, to paint the most comprehensive picture possible of DIBELS predictive power in prevention-oriented frameworks like Pennsylvania's Reading First Initiative and RTI, it is important to examine inconsistent achievement patterns to determine which students were "missed" by the system. 2006 first grade DIBELS results showed that 538 students who were expected to achieve "low risk" status on Spring oral reading fluency measures based on proficient Fall scores on segmenting/decoding tasks did not (PSF n=384, NWF n=154). In contrast, surprising improvements in performance were seen for 686 struggling first grade readers; those students moved from being in "at risk" performance categories in September to "low risk" categories in May.

These data suggest that the established cut-scores for both "low risk" and "at risk" performance levels on the Fall PSF resulted in an inappropriate balance of sensitivity and specificity. Under-identification of reading difficulties occurred for 15% of the Fall "low risk" cohort. Specifically, 760 students labeled "low risk" by DIBELS were misidentified. In the context of the 8,595 first grade students in Reading First Schools across the state who took the PSF subtest in the Fall, 760 doesn't seem like a significant number of bypassed children. Practically, however, if "at risk" children in Reading First are supposed to receive at least 120 minutes of additional intensive reading intervention per week each of those 760 students were each denied approximately 4,320 minutes of reading support during 36 weeks of the school year. This adds up to 3,283,200 minutes of reading intervention misappropriated in the PA Reading First first-grade classrooms studied. At the individual classroom level, the results suggest that in a typical first-grade classroom of 25 students, over 3 students would have been essentially misdiagnosed by PSF DIBELS results. Those false negative decisions weaken early intervention efforts by depriving truly at-risk students the intervention required to promote long-term reading success.

Over-identification of reading difficulties occurred for 26% of the Fall "at risk" cohort. Specifically, 336 students labeled "at risk" by DIBELS were misidentified. Again, stepping back from the large sample of first grade students studied overall, those results indicate that critical reading instructional time and intervention resources were spent on first grade students who may not have needed them, while the aforementioned group of false-negative students likely received no additional support. If DIBELS PSF cut-scores in the Fall of first grade had higher degrees of specificity, 1,451,521 minutes of reading intervention could have been reallocated to the population of students in more need of intensive instruction.

The detected high rates of sensitivity for the first grade DIBELS PSF measures corroborated with other recent research. In examining PSF prediction validity for kindergarteners, Hintze and colleagues (2003) found that concerning numbers of students were labeled "at risk" by DIBELS PSF measures who did not demonstrate low performance on other tests of phonological ability. Moreover, unpublished dissertation findings examining the predictive power of DIBELS LNF, PSF, and NWF subtests administered in the Fall to first graders showed that PSF achievement thresholds produced concerning numbers of false positives when predicting Spring DORF outcomes (Ryan, 2004).

Unfortunately, the new cut-points derived from the ROC Curve analyses results in the current research were more liberal than conservative. Essentially, to achieve a statistically appropriate balance between sensitivity and specificity resulting in an area under the Curve that exceeded .75, the PSF cut-scores for Fall "low risk" and "at risk" status were lowered. These changes would *increase* the number of true negatives and true positive predictions, meaning more children would be accurately labeled good readers and more children would be accurately labeled poor readers; with a "low risk" cut-score of 30 cspm on the Fall PSF (instead of 35) DIBELS would identify 17% more truly skilled readers. However, false negative rates would also *increase* by 2%. Therefore, rather than missing 3 children at risk for reading failure in a typical first grade classroom, 4 would be overlooked.

As far as reading intervention and resource allocation for struggling readers goes, decreasing the "at risk" cut-score to 8 sounds per minute (as recommended by the ROC analysis) would mean that schools could appropriately provide additional intensive reading support to 531 more readers than when using the author-recommended cut point of 10 cspm. Again, this positive result is coupled with some negative drawbacks. By increasing the pool of children

accurately identified as at risk for reading disabilities, the rate of inaccurate identification also increases. In this case, the group of first grade Reading First students who would have received unnecessary reading intervention last year would have increased by 40 children state-wide.

The established cut points for the NWF subtest were statistically appropriate. In fact, the area under the ROC curve for the cut point of 24 cspm for "low risk" performance exceeded both "fair (greater than or equal to .75) and "good" (greater than or equal to .8) requirements for diagnostic accuracy suggested by Swets et al. (1988). The same was true for the cut point of 13 cspm for "at risk" performance. However, again it is important to consider the implications of the resulting over and under-identification rates. While statistically reasonable, the results are practically concerning. While 3,553 first graders targeted for reading success by NWF guidelines became good readers, 816 did not. Those 816 false negative predictions represent students who needed intensive instructional support in reading but likely did not receive it because of DIBELS misidentification. In contrast, the 343 students who were inaccurately classified as "at risk" and were allotted extra minutes of reading instruction per week in first grade did not need the intervention.

Overall, valuable time was lost for 1,576 truly "at risk" students who were missed by both the PSF and NWF standards. If the purpose of DIBELS is to help 'catch' students before they fall too far behind in reading, accurate identification and intensive intervention must happen early. We know that children who fail to demonstrate strong reading skills in first grade generally remaining poor readers (Felton & Wood, 1992; Juel, 1988). So, if research says that identification early in first grade is critical, DIBELS results unfortunately delayed intervention for the equivalent of 63 entire first grade classrooms of students.

For second graders, achievement patterns suggested that DIBELS benchmarks goals on the DORF resulted in consistent classification of students overall. However, the diagnostic accuracy findings highlight some important drawbacks. These data indicated that the areas under the ROC Curves generated for the established 2nd grade cut-scores for both "low risk" and "at risk" performance levels on the Fall DORF exceeded requirements for both "excellent" (.92) and "good" (.82) diagnostic criteria (Swets, 2003). Under-identification of reading difficulties occurred for only 11% of the Fall "low risk" cohort. Specifically, 428 students labeled "low risk" by DIBELS were misidentified. In the broad context of the 7,925 second grade students studied, incorrectly grouping approximately 1/18 of the sample does not seem significant. However, considering previous research findings that show if students' reading difficulties are not detected until the end of second grade it may be too late for classroom instruction to have a significant impact on reading acquisition, missing 428 students becomes more serious (Chard & Kame'enui, 2000). Those students missed out on 1,848,960 minutes of reading intervention during second grade.

Over-identification of reading difficulties occurred for only 10% of the Fall "at risk" cohort. Two hundred sixteen students falsely identified as "at risk" were candidates for more focused reading instruction including more time, smaller student teacher ratios, and more frequent progress monitoring than their peers. Those provisions were misallocated for that population of students. If typical intensive intervention groups match 5-6 students per teacher then 36 teachers in Reading First schools could have been better utilized to meet the needs of truly at risk students.

Analyses of 2006 third grade data beyond the mere frequencies of students falling into each performance category throughout the year showed that the Fall DORF cut-points resulted in

the largest percentages of true negative and true positive decisions for predicting "low risk" and "at risk" outcomes of all grade-level assessments examined. For example, requiring children to read at least 77 words correctly per minute in September of third grade to be considered at "low risk" for reading difficulties at the end of third grade correctly identified 90% (2,894) of students across the state. Moreover, this cut-score produced a ROC Curve area of .96 indicating "excellent" sensitivity/specificity levels (Swets et al. 1988). Statistically speaking, misidentifying 450 students as "low risk" (i.e., students who later exhibited significant reading difficulties) is a reasonable error rate. Practically speaking, educators likely would not agree. Missing an opportunity to positively impact reading growth for one out of every seven students (14%) in a classroom seems professionally inappropriate. The educational fate of these 450 students seems particularly bleak considering that patterns of reading failure established in the first three years of schooling are very difficult to change without explicit, intense, research-based reading intervention (NRC, 1998).

DIBELS authors propose that it is highly predictive of performance on high-stakes outcome measures (Good et al., 2001). In the 2001 validation study, Good and colleagues reported that 96% of the third grade students who achieved "low risk" status on DIBELS oral reading fluency tasks in the Spring also achieved or exceeded proficiency standards on the Oregon State Assessment. Results were not as dramatic for Pennsylvania's third graders in Reading First on the PSSA. The analysis of 2006 achievement data showed that 77% of third grade students who were scored in the "low risk" range on DIBELS Spring DORF also scored in the Proficient or Advanced range on the Reading subtest of the PSSA. Additional analyses indicated DIBELS hit rates for true negative vs. false negative identification (i.e. truly good readers vs. not good readers) were statistically appropriate with an area under the ROC Curve

exceeding minimal acceptability requirements (.83). 2,268 good readers made adequate progress in the general reading curriculum to not only meet the high-stakes reading goals of the PSSA but also were statistically likely to have long-term reading success. Unfortunately, 500 misidentified students experienced reading failure instead of success. That means that 3-4 students in each of the 145 Reading First schools studied who would have been expected to pass the PSSA based on high DIBELS scores would have failed. Imagine a teacher's dismay if four students in her class performed at Basic or Below Basic levels on the PSSA when DIBELS results suggested they would pass.

DIBELS hit rates for true positive vs. false positive identification (i.e. truly at risk readers vs. not at risk) were statistically inappropriate with an area under the ROC Curve falling short of the standard (.73). Using the author recommended "at risk" cut point of 80 correct words per minute, DIBELS incorrectly identified 477 third graders. Performance expectations for those third graders were probably low based on deficient DIBELS performance even though that cohort of students ultimately achieved success on grade-level PSSA reading tasks. The establishment of a new DORF cut-point would do little to remedy disproportionate true positive vs. true negative classifications. In fact, to achieve the statistical standard of a "fair" balance of sensitivity and specificity (greater than .75) true negative decisions would increase only by 4% whereas, false negative decisions would increase by 3%. Ultimately, 544 students would have received unnecessary reading support. The analysis did not produce any cut-scores for the third grade DIBELS DORF subtest that would reduce the number of false positive determinations.

5.1.2 DIBELS Prediction of Long-term Achievement Outcomes

Good et al. (2001) recommended that future research focus on examining the utility of DIBELS first grade measures to predict long-term reading performance, specifically on high-stakes reading assessments. Results from the current study showed that students' first grade performance on the Winter DORF subtest was more predictive of end-of-third grade achievement on the DORF than first grade performance on the Fall or Winter NWF or PSF subtests. But, despite significant p values, first grade fluency rates explained only 15% of the variance in students' third grade fluency scores; 85% of the remaining variance was left unexplained. The combination of all Level 2 variables (i.e., 1st Grade DIBELS Fall/Winter scores; minority status; SES; school) resulted in a moderately strong correlation with long-term third grade outcomes on the DORF. Ultimately, one-quarter of the variability in end-of-third grade DIBELS scores was explained by the combination of the six variables entered into the HLM model. The significance of these predictors should be interpreted cautiously however and may be inflated by the high power of the large sample size. Therefore, to determine the practical significance of each predictor, interpretation focused on the Empirical Bayes Estimates produced in the analysis. These test statistics are unique coefficients produced by HLM growth analysis that express the amount each variable predicts outcomes by contributing to students' growth over time. These estimates are interpreted similarly to correlation coefficients; the strength of the prediction increases as the value approaches 1 (Raudenbush & Bryk, 2002).

The results suggested that students' achievement on the DORF in January of first grade were moderately predictive of their third grade achievement. The PSF subtests and NWF subtests were far less predictive. In fact, students' race, poverty level, and the school they attended predicted their reading progress more than their ability to effectively segment phonemes

on the DIBELS PSF subtest in September of first grade. NWF achievement was more predictive than students' economic status but only impacted growth as much as race and school. These results compounded by the false negative and false positive identification rates discussed earlier raise important questions about DIBELS effectiveness as an "early" indicator of reading skill. If the subtests administered in the Fall of first grade are not predictive of future reading achievement, DIBELS does not achieve its primary purpose.

5.1.3 DIBELS Prediction of Long-term Achievement Outcomes on the PSSA

Similar results occurred for the HLM analysis of students' first grade DIBELS to third grade PSSA achievement. When examining the predictive relationship between students' early first grade DIBELS scores and their eventual 3rd grade achievement on the PSSA, we see that there was significant movement in and out of risk groups across the three years of schooling. One would hope that this movement would reflect student reading growth by changing from "at risk" on DIBELS subtests in first grade to "proficient" on the PSSA. Unfortunately, nearly equal proportions of students moved from "low risk" on DIBELS to "not proficient" on the PSSA as did their more successful peers who demonstrated reading growth. Why such inconsistency of achievement? The results of the HLM analysis suggest that the DIBELS measures administered in first grade were generally not predictive of third grade reading achievement for students in these Reading First schools. In fact, the three reading subtests designed to be indicators of students overall reading "well being" explained only 18% of the variability in PSSA Reading scores. Once again, first grade Winter DORF was most highly correlated with PSSA achievement (.40). Furthermore, examination of the Empirical Bayes Estimates shows that this subtest was also the strongest predictor of 3rd grade reading scores. NWF achievement and

students' economic status were equally predictive; whereas the PSF was not found to be a significant predictor at all. The minimal impact of students' performance on the PSF subtest on later PSSA proficiency parallels other prediction studies. While researchers agree that phonemic awareness is a critical prerequisite skill for competent reading, classification studies reveal that over-reliance on phonological tasks results in high proportions of false positives for identifying reading disabilities and overall inaccurate prediction of reading skill (Scarborough, 1998; Speece & Case 2001; Speece, Mills, Ritchey & Hillman, 2003).

5.1.4 DIBELS Prediction of Special Education Eligibility

The correlation and regression coefficients examining the relationship between students' risk classification on first grade DIBELS, students' race, and third grade special education status were highly concerning but not surprising given earlier findings. Students' first grade oral reading fluency rates served as the best reading predictors of disability status. However, in keeping with the patterns of minority overrepresentation identified by the NRC (2001), non-white students were more likely to be identified for special education than their peers. The significant relationships between students' minority status and poverty level with disability status may also be related to the unique demographics of Reading First schools.

DIBELS classification of first grade students as "at risk" for reading difficulties appears relatively inaccurate given the high numbers of students not ending up in special education by third grade. On the one hand, the limited number of students identified as having reading disabilities by May of 2006 may be attributed to the effectiveness of Reading First; students identified as "at risk" early in their school career received enough reading intervention and support to eventually successfully "respond" to instruction. Future research should examine

longitudinal referral and identification trends for "at risk" students in PA Reading First Schools to determine the impact of its Three Tier intervention framework on special education eligibility determinations. On the other hand, the data suggests that the DIBELS subtests administered in first grade produced drastic numbers of false negative and false positive predictions for disabilities. In fact, over 400 first grade students who were not flagged for intensive interventions based on PSF, NWF, or DORF results in first grade were eligible for special education by third grade. Once again, when related to a classroom of first graders, DIBELS inaccurate classification rates would overlook approximately 3 students per class. The average number of first grade classrooms in the Reading First schools studied was 3 per building. Suggesting that 9 students in each Reading First school were victims of the "wait to fail" system because they were missed by DIBELS in first grade is a clear extrapolation; nonetheless, these diagnostic (in)accuracy results raise numerous red flags.

The high rates of false positive predictions for disability status are also concerning. Over-identification of first graders occurred for approximately 1,000 students based on Fall DIBELS classifications of "at risk". Pressley and colleagues (2006) caution DIBELS users that the financial strain put on school resources by intervening with the "wrong" population of students is great—exceeding $90 per student per year.

Logistic regression results strengthen the case for using DIBELS Winter DORF results as indicators instead of Fall PSF and NWF subtests. The findings confirm decades of prior research reporting that oral reading fluency serves as the best predictor of future reading competence. However, if striving to identify which variables are most predictive of special education, race and poverty are the unfortunate favorites.

5.1.5 Limitations

Given that the study examined Pennsylvania Reading First data, the samples of students studied were drawn from schools that were culturally diverse, low-achieving, and economically depressed. Overall, the results should be generalized with caution. Moreover, Reading First favors a three-tiered system of instruction/intervention in which students flagged as at-risk for reading failure should receive more intense, more focused instruction geared to improve reading outcomes. The present study did not measure or control for the presence, nature, frequency, or intensity of instructional intervention for students labeled "at risk" by DIBELS. Therefore, future research should explore the diagnostic accuracy of DIBELS "at risk" cut points in schools with tiered intervention systems and schools without tiered intervention systems to determine the whether movement from "at risk" to "low risk" categories results from the inaccuracy of DIBELS predictions or from the effectiveness the applied reading intervention.

An additional limitation of the study involves excluding incomplete records from the analyses. By examining growth patterns for students with complete data sets only, important nuances of achievement trends that may be related to attrition, attendance, migration etc. were not considered. Additional analyses of trends in the missing data are important. Finally, it is important to note two limitations in the analyses of special education eligibility status. First, identification was based on students' primary disability classification on the third grade PSSA. Therefore, students' exact "entry" dates into special education were unknown. Second, because the disability categories used in the analyses may have included children who did not receive special education services for reading, broad generalization of the findings is not appropriate. Future research should focus on disaggregating the results by primary disability category and type of support to better reflect DIBELS prediction of specific reading disabilities.

5.1.6 Future Research

The findings of the present study point to the need for more research examining DIBELS predictive strength in early intervention and RTI systems to better operationalize educational decision-making for our neediest students. Exploration of DIBELS' prediction of students' achievement on norm-referenced, standardized tests other than state-wide achievement tests is important. Examining the relationship between students' first and second grade DIBELS performance, Terra Nova achievement, and later PSSA achievement would not only shed some light on DIBELS predictive validity for a more widely-used test, but may also highlight important strengths and limitations of using the PSSA as an absolute measure of reading achievement.

Analyzing the attributes of the students who were under-identified (i.e. false negatives) by DIBELS would also be interesting. If the RTI approach focuses on moving "at risk" students through varying levels of intervention to determine whether they have a true disability and are in need of special education, then correctly identifying the initial "at risk" population is critical. The situation for students who are false positives who will eventually respond to intervention and move out of the system is less grave than it is for their false negative counterparts who were completely missed to begin with. If research can help identify the characteristics and skill profiles of those "false-negative" students, we might be able to catch them (despite DIBELS' lack of specificity) before they fall too far behind. Moreover, considering that the ROC analyses in this study did not produce any new cut points that resulted in fewer false negative predictions, future research might examine the impact of using various combinations of DIBELS results to

determine risk status in an effort to increase specificity. For example, what would the area under the ROC Curve and the sensitivity v. specificity indices be for a combined battery of DIBELS subtests of Fall NWF, Fall PSF, and Winter DORF?

Above all, it is important that research continue to explore the most effective and most appropriate uses of this assessment in a current reform climate, which relies on DIBELS results to inform instructional practices, impact early intervention decisions, and inform educational policy.

5.1.7 Summary

The primary goal of this research was to add to the slowly growing research base examining DIBELS' effectiveness in informing educational decisions and identifying children considered to be at-risk for reading failure. The overall limited predictive value of DIBELS on students' long-term reading achievement raises important concerns about over-reliance on DIBELS in an early intervention framework like Pennsylvania's Reading First. If the goal of federal initiatives like Reading First is to increase our country's literate population, it is imperative that no children are actually "left behind" (especially the poor readers). Moreover, the effectiveness of a school-wide educational decision making system such as Response-to-Intervention (RTI) directly relates to the accuracy of the measure used to identify students at-risk for reading problems because of the seriousness of the decisions made based on test performance such as group placement, intervention intensity, and referral for special education. In a recent publication Kenneth Goodman and colleagues (2006) hypothesize that the political and professional pressure placed on schools to meet accountability standards and reduce special

education referrals at all costs clouds our ability to carefully scrutinize the strengths and weaknesses of this assessment.

It is true that no single assessment is an absolute measure of students' reading skill and DIBELS authors state that it is not a diagnostic tool and should be used only as a "thermometer" of reading health. However, they also suggest that when used as recommended, DIBELS results and benchmark classifications can be used to effectively evaluate individual student development, predict later reading proficiency, and aid in early identification. To achieve that, we may need a better instrument. Personally, I would rather use a thermometer that consistently and accurately indicates whether or not I had a fever before I paid for unnecessary treatment. Even more importantly, I would rather use a thermometer that consistently and accurately indicates whether or not I had a fever before an undetected condition becomes serious and incurable.

REFERENCES

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

Beck, I.L. & Juel, C. (1995). The role of decoding in learning to read. *American Educator*, 19(2), 21-25.

Blachman, B.A. (1997). Early Intervention: A cautionary tale. In B. A. Blachman (Ed)., *Foundations of reading acquisition and dyslexia: Implications for early intervention* (pp. 409-430). Mahwah, NJ: Lawrence Erlbaum Associates.

Blachman, B.A. (2000). Phonological Awareness. In M.L. Thomas, P.B. Mosenthal, P.D. Pearson, & R. Barr (Eds.), *Handbook of Reading Research* (Vol. III, pp. 483-502.) Mahwah, NJ: Lawrence Erlbaum Associates.

Bloom, B.S. (1976). *Human Characteristics and School Learning*. New York: McGraw Hill.

Braden, J.P. (2002). Best practices for school psychologists in educational accountability: High stakes testing and educational reform. In A. Thomas, & J/ Grimes (Eds.), *Best Practices in School Psychology IV* (pp. 310-319). Bethesda, MD: National Association of School Psychologists.

Bradley, L., & Bryant, P. E. (1983). Categorizing sounds and learning to read: A causal connection. *Nature*, 301, 419-421.

Bryk, A.S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage Publications.

Chall, J.S. (1989). Learning to read: The great debate 20 years later: A response to "Debunking the great phonics myth." *Phi Delta Kappan*, 70, 520-538.

Chard, D.J., & Kameenui, E.J. (2000). Struggling first-grade readers: The frequency and progress of their reading. *The Journal of Special Education,* 34, 28-38.

Center for the Improvement of Early Reading Achievement. (2001). Put reading first: The research building blocks for teaching children to read. The Partnership for Reading: National Institute for Literacy; National Institute of Child Health and Human Development; and U.S. Department of Education.

Coyne, M. D., Kame'enui, E.J., & Simmons, D.C. (2001). Prevention and intervention in beginning reading: Two complex systems. *Learning Disabilities Research and Practice,* 16 (2), 62-73.

Deno, S. L. (1989). Curriculum-based measurement and special education services: A fundamental and direct relationship. In M. R. Shinn (Ed.), Curriculum Based Measurement: Assessing Special Children (pp.1-17). New York: The Guilford Press.

Ehri, L.C. (1991). Learning to read and spell words. In Rieben, & C.A. Perfetti (Eds.), *Learning to Read: Basic research and its implications* (pp. 57-73). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ehri, L.C. (1992). Reconceptualizing the development of sight word reading and its relationship to recoding. In P,B, Gough, L.C. Ehri, & R. Treiman (Eds.), *Reading Acquisiton* (pp. 107-143). Hillsdale, NJ: Lawrence Erlbaum Associates.

Felton, R. H., & Wood, F. B. (1992). A reading level match study of nonword reading skills in poor readers with average IQ. *Journal of Learning Disabilities,* 25, 318- 326.

Fletcher, J.M., Morris, R., Lyon, G.R., Stuebing, K.K., Shaywitz, S.E., Shankweiler, D.P., Katz, L., & Shaywitz, B.A. (1997). Subtypes of dyslexia: An old problem revisited. In B. A. Blachman (Ed)., *Foundations of reading acquisition and dyslexia: Implications for early intervention* (pp. 95-114). Mahwah, NJ: Lawrence Erlbaum Associates.

Fuchs, L. S., & Fuchs, D. (2000). Monitoring students Progress toward the development of reading competency. *School Psychology Review,* 38 (4), 650-671.

Fuchs, L. S., & Fuchs, D. (1986). Curriculum-based assessment of progress toward long- and short-term goals. *The Journal of Special Education,* 20, 69-82.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989). Computers and curriculum-based measurement: Effects of teacher feedback systems. *School Psychology Review,* 18 (1), 112-125.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Allinder, R. M. (1991). The reliability and validity of skills analysis within curriculum-based measurement. *Diagnostique,* 14, 03-221.

Fuchs, L.S., Fuchs, D., Hosp, M. K., & Jenkins, J.R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, 5, 239-256.

Good, R.H., III & Kaminski, R.A. (1996). Assessment for instructional decisions: Toward a proactive/prevention model of decision-making for early literacy skills. *School Psychology Quarterly*, 11, 326-336.

Good, R.H. III, & Kaminski, R.A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills (6th ed.)*. Eugene, OR: Institute for the Development of Education Achievement. Available: http://dibels.uoregon.edu

Good, R.H. III, Kaminski, R.A., & Kame'enui, E.J. (2001) Using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an outcomes-driven model. *OSSC Bulletin*, 44(1), 1-24.

Good, R.H., III, 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, 259-290.

Good, R.H. III, Simmons, D., Kame'enui, E.J., Kaminski, & Wallin, J. (2002). *Summary of decision rules for intensive, strategic, and benchmark instructional recommendations in kindergarten through third grade* (Technical Report No. 11). Eugene, OR: University of Oregon.

Goodman, K.S., Flurkey, A., Kato, T., Kamii, C., Manning, M., Seay, S., Thome, C., Tierney, R.J., Wilde, S. (2006). *The Truth About DIBELS: What it is, What it Does.* Portsmouth, NH: Heinemann.

Grimm, L.G. & Yarnold, P.R. (Eds.). (1995). *Reading and Understanding Multivariate Statistics.* Washington D.C.: American Psychological Association

Hintze, J., Ryan, A., Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Processing. *School Psychology Review* 12 (4), 541-556.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, 80, 437-447.

Kame'enui, E.J., & Carnine, D.W. (Eds.). (1998). *Effective teaching strategies that accommodate diverse learners.* Columbus, OH: Merrill.

Kaminski, R. A., Good III, R. H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review*, 25 (2), 215-228.

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

Lundberg, I. (1991). Phonemic awareness can be developed without reading instruction. In S.A. Brady, & D.P. Shankweiler (Eds.), *Phonological processes in literacy: A tribute to Isabelle Y. Liberman* (pp. 47-53). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lyon, G. R, (1999). In celebration of science in the study of reading development, reading difficulties and reading instruction: The NICHD perspective. *Issues in Education,* 5 (1), 85-116.

Meyer, M.S., & Felton, R.H. (1999). Repeated reading to enhance fluency: Old approaches and new directions. *Annals of Dyslexia*, 49, 283-206.

National Assessment of Educational Progress. (2005). http://nces.ed.gov/nationsreportcard

National Institute of Child Health and Human Development (NICHD). (2000). Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction; Reports of the subgroups. (NIH Publication No 00-4754). Washington, DC: U. S.

National Reading Panel (2000). Teaching children to read: An evidence-based assessment of scientific research literature on reading and its implication for reading instruction. Washington, DC: National Institute of Child Health and Human Development.

National Research Council (1998). Preventing reading difficulties in young children. Washington, DC: National Academy Press.

O'Connor, R.E., & Jenkins, J.R. (1999). Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading,* 3 (2), 159-197.

Perfetti, C.A. (1985). *Reading ability*. New York: Oxford University Press.

Perfetti, C.A., Beck, I., Bell, L., & Hughes, C. (1987). Phonemic knowledge and learning to read are reciprocal: A longitudinal study of first grade children. In K. Stanovich (Ed.), Children's reading and development of phonological awareness. Merril Palmer Quarterly, 33 (3), 283-320.

Pressley, M., Hilden, M., Shankland, R. (2006). *An Evaluation of End-Grade-3 Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Speed reading without comprehension, predicting little.* Unpublished manuscript.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models, applications and data analysis methods (2nd ed.)*. Thousand Oaks, CA: Sage Publications.

Ravitch, D. (1999). Student performance: The national agenda in education. In M. Kanstroom, & C.E. Finn (Eds.), *New directions: Federal education policy in the twenty-first century.* Washington, DC: Thomas B. Fordham Foundation & The Manhattan Policy Institute.

Ryan, A. (2004). *Diagnostic Accuracy of the Dynamic Indicators of Basic Early Literacy Skills in the Prediction of First Grade Oral Reading Fluency* Unpublished dissertation, University of Massachusetts Amherst

Salvia, J., & Ysseldyke, J.E. (1998). *Assessment* (7th ed.). Boston, MA: Houghton Mifflin.

Schreiber, P.A. (1980). On the acquisition of reading fluency. *Journal of Reading Behavior,* 12, 177-186.

Share, D.L., Jorm, A.F., Maclean, R., & Matthews, R. (1984). Sources of individual differences in reading achievement. *Journal of Educational Psychology*, 76 (6), 1309-1324.

Share, D.L., & Stanovich, K.E. (1995). Cognitive processes in early reading development: Accommodating individual differences into a model of acquisition. *Issues in Education*, 1, 1-57.

Shaywitz, S.E. (1996). Dyslexia. *Scientific American,* 275 (5), 98-104.

Speece, D.L., & Case, L.P. (2001.) Classification in Context: An alternative approach to Identifying early reading disability. *Journal of Educational Psychology*, 93, 735-749.

Speece, D.L, Case, L.P, Molloy, D.E. (2003) Responsiveness to General Education Instruction as the First Gate to Learning Disabilities Identification. *Learning Disabilities Research & Practice,* 18 (3), 147-156.

Stanovich, K.E. (1992). Speculations on the causes and consequences of individual differences in early reading acquisition. In P,B, Gough, L.C. Ehri, & R. Treiman (Eds.), *Reading Acquisiton* (pp 307-342). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stanovich, K., & Stanovich, P. (1995). How research might inform the debate about early reading acquisition. *Journal of Research in Reading,* 18, 87-105.

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240, 1285-1293.

Swets, J.A., Dawes, R.M. & Monahan, J. (2003) Psychological science can improve diagnostic decision. *Psychological Science in the Public Interest*, 1, 1-26.

Tabachnick, B.G. & Fidell, L.S. (1996). *Using Multivariate Statistics*. NY: HarperCollins

Tatano-Beck, C., & Gable, R.K. (2001). *The area under a ROC curve*. http://gim.unmc.edu/dxtests/roc3htm

Torgeson, J.K. (1998). Catch them before the fall: Identification and assessment to prevent reading failure in young children. *American Educator*, 22 (1), 32-39.

Torgeson, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem. *Learning Disabilities Research and Practice,* 15, 7 (4), 303-323.

Torgeson, J. K., Alexander, A. W., Wagner, R. K., Raschotte, C. A., 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, 78.

Torgeson, J.K., Wagner, R.K., Rashotte, C.A. (1997). Prevention and Remediation of Severe Reading Disabilities: Keeping the End In Mind. *Scientific Studies of Reading*, 1 (3), 217-234.

U.S. Department of Education web site. Executive Summary- No Child Left Behind. Retrieved April 3, 2006 from http://www.ed.gov/nclb/overview/intro/execsumm.html

U.S. Department of Education web site. Introduction- Reading First. Retrieved April 3, 2006 from http://www.ed.gov/programs/readingfirst/index.html

U.S. Department of Education web site. Proposed Regulations- Individuals with Disabilities Education Improvement Act of 2004 (IDEA). Retrieved April 3, 2006 from http://www.ed.gov/policy/speced/guid/idea/idea2004.html

Valencia, S.W., & Buly, M.R. (2004) Behind Test Scores: What Struggling Readers Really Need. *The Reading Teacher*, 57 (6), 520-531.

Vaughn, S., Linan-Thompson, S., & Hickman, P. (2003). Response to instruction as a means of identifying students with reading/learning disabilities. *Exceptional Children*, 69, 391-409.

Vellutino, F.R., Scanlon, D.M., Sipay, E.R., Small, S.G., Pratt, A., Chen, R.S., & Denckla, M.B. (1996). Cognitive profiles of difficult to remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology,* 88, 601-638.

Wagner, R.K. & Torgeson, J.K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin,* 101, 192-212.

Wolf, M., & Bowers, P. (1999). The "double-deficit hypothesis" for the developmental dyslexias. *Journal of Educational Psychology*, 91 (3), 1-24.

Yopp, H. K. (1988). The validity and reliability of phonemic awareness tests. Reading Research Quarterly, 23, 159-177.