5-21-2008

Using Reading Screening Measures to Create Risk Indicators for Student Performance on the New York State English Language Arts Examination

Kimberly Davis

Using Reading Screening Measures to Create Risk Indicators for Student Performance on the New York State English Language Arts Examination

Graduate Thesis

Submitted to the Faculty

Of the School Psychology Department

College of Liberal Arts
ROCHESTER INSTITUTE OF TECHNOLOGY

By

Kimberly Davis

In Partial Fulfillment of the Requirements
For the Degree of
Master of Science and
Advanced Graduate Certificate

Rochester, New York May 21, 2008

Approved: _____
(committee chair)

_____
(committee member)

RIT
School Psychology Program
Permission to Reproduce Thesis

PERMISSION GRANTED
Title of thesis Using Reading Screening Measures to Create Risk Indicators for Student Performance on the New York State English Language Arts Examination

Kimberly M. Davis hereby **grant permission** to the Wallace Memorial Library of the Rochester Institute of Technology to reproduce my thesis in whole or in part. Any reproduction will not be for commercial use or profit.

Date: 6/13/08 **Signature of Author:** _____

PERMISSION FROM AUTHOR REQUIRED
Title of thesis _____
_____
_____

I _____ **prefer to be contacted** each time a request for reproduction is made. I can be reached at the following address:

_____
_____
_____

PHONE: _____

Date: _____ **Signature of Author:** _____

PERMISSION DENIED
TITLE OF THESIS _____
_____
_____

I _____ hereby **deny** permission to the Wallace Memorial Library of the Rochester Institute of Technology to reproduce my thesis in whole or in part.

Date: _____ **Signature of Author:** _____

Running head: USING READING SCREENING MEASURES

Using Reading Screening Measures to Create Risk Indicators for Student Performance on the New York State English Language Arts Examination

Kimberly M. Davis

Rochester Institute of Technology

Abstract

The purpose of this study was to (a) determine whether the Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency (DORF) measure and the Developmental Reading Assessment (DRA) will accurately predict student failure on statewide assessments of reading performance, and (b) establish risk indicators for both DORF and the DRA that are predictive of student failure on a statewide reading assessment. One hundred ninety-five second grade students were administered DORF probes during the fall, winter, and spring and the DRA during the fall and spring. They were then administered the New York State English Language Arts Examination (NYS ELA) during January of their third grade year. Patterns of correlations between the two potential screening measures and the NYS ELA were examined. Risk indicators for predicting student performance on the NYS ELA were established using Receiver Operator Characteristic (ROC) curve analysis. Results indicated both DORF and the DRA were moderately effective at predicting student performance on the ELA during the following school year. Comparisons between risk indicators established in the present study and previously-established district benchmarks were made.

CHAPTER I

Statement of the Problem

The importance of literacy in modern society cannot be overstated. In the United States today, the ability to read is essential because it provides access to learning, politics, and economic success (Brandt, 2001). In order to be successful workers in today's society, it is imperative that high school graduates be able to read complex material. In essence, 100 percent literacy rates are expected of today's youth. However, despite the importance of reading today, many American children cannot read by the time they leave high school (Burns, Griffin, & Snow, 1999). The 2007 National Assessment of Educational Progress (NAEP) reported on the percentages of students across the nation performing within expected levels in reading. Results indicated 34 percent of fourth grade students were reading below the basic level of proficiency. In other words, 34 percent of students were not performing at a level in reading that would enable them to complete the work assigned in that grade (Lee, Grigg, & Donahue, 2007).

Illiteracy affects children from all social categories, ethnicities, and cultures; however, it is most prevalent in children from low socioeconomic backgrounds, minority cultures, and children whose native language is not English (Burns et al., 1999). Large discrepancies have been noted regarding differences in student reading abilities in poverty-stricken areas. The 2007 NAEP report noted that 50 percent of economically disadvantaged students identified by their eligibility for free or reduced-cost lunch scored below the basic achievement standard set by NAEP as opposed to 21 percent of students not eligible for free or reduced-cost lunch (Lee, Grigg, & Donahue, 2007). In addition, children from racial or ethnic minority groups were found to perform below the basic achievement standard set by NAEP more often than Caucasian students. Fifty-four percent of African American students, 51 percent of Hispanic students, 24

percent of students of Asian/Pacific Island descent, and 49 percent of American Indian/ Alaskan Native students scored below the basic achievement standard set by NAEP as opposed to 23 percent of White students (Lee, Grigg, & Donahue, 2007).

The effect of illiteracy on American society is portrayed through numerous statistics. For example, illiteracy affects 75 percent of unemployed individuals, 85 percent of juveniles who appear in court, and 60 percent of prison inmates (Adams, 1990). Recent technological advances have further increased the demand placed on individuals to be able to read in order to function effectively in modern society (Adams, 1990).

Children exhibiting reading difficulties early in their schooling may continue to experience difficulty with reading throughout their educational careers. For example, children exhibiting reading difficulties in first grade are highly likely to continue to have difficulties with reading in fourth grade (Juel, 1988). Furthermore, research suggests good readers read many more books than poor readers. This additional reading experience for good readers is likely part of the reason that they are apt to remain good readers over time while poor readers are not likely to become good readers. These findings indicate that early intervention with young struggling readers is necessary to ensure a pattern of poor reading performance does not follow these children throughout their school careers (Juel, 1988).

In recent years, increased support has been established for the theory that reading performance is highly influenced by performance in areas of early literacy (National Reading Panel, 2000). In 1997, the United States Congress commissioned the National Reading Panel (NRP) to assess the large base of research regarding the acquisition of early literacy skills and submit a formal report to Congress in February of 1999 (NRP, 2000). Stringent criteria were involved in the selection of research studies in order to provide the most current, in-depth

information regarding literacy development and the teaching of early literacy skills (NRP, 2000). The Panel's report discusses findings related to five "big ideas" or components of reading which are: phonemic awareness, alphabetic principle, vocabulary, accuracy and fluency, and comprehension. The Panel also highlighted the importance of fluency as one of the main components needed for reading comprehension (NRP, 2000, p.11). However, research suggests fluency tends to be overlooked in the classroom. In order to improve reading fluency skills, students must practice reading (NRP, 2000).

Attempts at Increasing National Reading Attainment

In order to address the issue of reading attainment in schools in the United States, the government enacted legislation in 2001 requiring certain standards be put into place for reading instruction and assessment. The 2001 reauthorization of the *Elementary and Secondary Education Act of 1965* involved a new component known as *No Child Left Behind* (NCLB) (U.S. Department of Education, 2004). This legislation has many parts; however, one main purpose of the legislation is to close the achievement gap among minority and non-minority students by providing a more inclusive and fair education for all children in the United States (U.S. Department of Education, 2004).

One part of the law includes a plan set forth by the national government that asks the states to set certain standards that school districts must meet in order to receive financial support. The NCLB legislation mandates that third through eighth grade students reach proficient levels of performance in core subjects by the 2013-2014 school year (NCLB, 2001). Until this date, schools must show that their students are making Adequate Yearly Progress (AYP) such that the discrepancy between the school's performance and a universal performance criterion is decreased within an allotted time frame. AYP is measured through the use of high-stakes tests of

achievement (Fuchs & Fuchs, 2004). For reading assessments in particular, the states choose the test that is given; however, the components of the assessment must be aligned with the reading and language arts standards delineated by the NCLB legislation (NCLB, 2001). Examples of these types of assessments are the New York State English and Language Arts Examination, the Oregon Statewide Assessment, and the Washington Assessment of Student Learning. Student performance on these norm-referenced tests is meant to represent the quality of education provided by the school. Therefore, the results of high-stakes testing have become extremely important to districts since the NCLB legislation went into effect in 2001 (Hintze & Silberlitt, 2005).

One program, Reading First, was developed in 2001 as a result of the NCLB legislation (U.S. Department of Education, 2002). This program provides financial assistance to schools to facilitate the implementation of scientifically-based reading instruction policies for students in kindergarten through third grade. Funding for this program is focused on schools and districts where a substantial portion of students are reading below grade level or are living in low-income homes. The goal of the program is that students will be competent readers by the end of third grade (U.S. Department of Education, 2002).

The Importance of Monitoring Student Achievement in Schools

The deleterious consequences of low literacy skills are both well-documented and broad. Therefore, an appropriate goal seems to be that of altering these negative outcomes and ensuring adequate literacy skills for all children. Formative progress-monitoring systems can provide data that is not available from summative academic assessments. These systems can provide data that is sensitive enough to inform teachers regarding the exact needs of individual students. Teachers can then design and focus instructional activities appropriately. Furthermore, monitoring systems

can be used to identify students in need of additional support earlier than more traditional practices of waiting for a child to be unsuccessful before providing additional support. Monitoring also provides the concrete information needed to identify children who may need additional support (Sloat, Beswick, & Willms, 2007).

Evaluation of Student Progress through Curriculum-Based Measurement

Curriculum-based measurement (CBM) was first suggested as a method of monitoring student achievement in the mid 1980s when consensus on how to monitor achievement did not exist (Deno, 1985). CBM is a tool that can be used to assess many different academic skill areas (e.g., mathematics, reading, spelling). According to Deno (1985), three components must be present in a measurement tool in order for it to be used effectively for assessment of student progress over time. These components include reliability and validity of the measure, simplicity and efficiency of the measure, and cost effectiveness of the measure. CBM is a measurement tool that satisfies these three criteria.

CBM procedures are sensitive to growth and enable even small changes in progress to be noted. Progress-monitoring data can be obtained in a time-efficient manner in order to enable teachers to make data-based decisions on ways to modify instruction to fit the needs of their students. In addition to monitoring student growth in overall reading development, teachers can use the information gathered from curriculum-based measures of reading (R-CBM) to analyze the types of errors students are making (Fuchs & Fuchs, 1999). Analysis of phonetic errors can provide teachers with information that can inform instruction in decoding skills. Moreover, due to the ability of CBM to be used repeatedly over time, many data points can be gathered to show a child's progress over time in comparison to same age peers.

One assessment system that is widely utilized to assess reading progress is the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), which was developed by researchers at the University of Oregon (Good, Kaminski, Simmons, & Kameenui, 2001). DIBELS assesses early literacy skills, including oral reading fluency, through the use of a series of short tests based upon CBM procedures. DIBELS is composed of seven tests including Initial Sound Fluency, Phoneme Segmentation Fluency, Nonsense Word Fluency, Letter Naming Fluency, Oral Reading Fluency, Retell Fluency, and Word Use Fluency. Oral Reading Fluency measures are used to assess students with a series of one-minute probes. Students are asked to read a short passage aloud while the examiner marks the number of words read correctly and the number of errors made in the one minute time period. Students are administered three of these probes and the median number of words read correctly and errors for the three probes is recorded as the student's Oral Reading Fluency score. That score can then be compared to benchmarks established from district or nationwide administration of the DIBELS measures to identify how a particular student is performing relative to other children in the same grade. The results of DIBELS assessment can also be used to track growth toward desired academic outcomes (Good, Kaminski, Simmons, & Kame'enui, 2001).

Over the past several years, an additional use of curriculum-based measures has been determined. Today, curriculum-based measures are being used as predictors of student performance on high-stakes achievement tests (Deno, 2003). Recent research has focused on correlating performance on curriculum-based measures with performance on high-stakes tests of student achievement (e.g., Hintze & Silberglitt, 2005; McGlinchey & Hixson, 2004). Additional research (e.g., Good, Simmons, & Kameenui, 2001) has attempted to provide benchmarks indicating levels of performance on curriculum-based measures that can be used to predict

performance on high-stakes assessments. The purpose of a benchmark goal is to identify a certain level of performance that is indicative of likely success on some specified outcome measure (Good, et al.). Benchmarks are established by combining a certain level of skill development with the time period in which that skill should be achieved. Ideally, students would be assessed using a particular screening system in order to determine which students are not meeting benchmark goals. These students would then be provided with additional support prior to the high-stakes assessment. Progress toward the benchmark goal would be monitored while intervention support was being provided (Good, Simmons, & Kameenui, 2001).

Another tool for assessing student reading progress in schools is the Developmental Reading Assessment (DRA). Developed in the Upper Arlington School District, the DRA is a widely used instrument for measuring reading achievement. According to the publisher of the DRA, it is used in more than 30,000 classrooms across the United States (Pearson Learning Group, 2003). The DRA measures three different components of reading including engagement, fluency and accuracy, and comprehension. The DRA is meant to be administered and interpreted by teachers. Each student is asked to read a series of short stories and answer questions related to those stories. Teachers then score the student's responses and arrive at a "level" indicative of the reading abilities of that student. Results for individual students can then be compared to identified standards for a particular grade level and those students in need of additional reading support can be identified (Beaver, 2004).

The usefulness of DORF and the DRA as measures that can provide valuable information about student performance in reading has been established. In addition, several studies have investigated the added use of DORF as a measure that can provide predictive information relating to outcome measures of student reading achievement. However, no research has been

conducted to date in which the DRA is utilized as a predictive tool for student performance on an outcome measure such as a high stakes test of reading achievement. More information is needed regarding the usefulness of both these measures as predictive tools.

Purpose of the Study

The present study will replicate and extend the work of Good and colleagues (2001) by determining the appropriateness of two commonly-used screening measures, DIBELS Oral Reading Fluency (DORF) and the DRA, in predicting student performance on a high-stakes reading assessment (Good, Simmons, & Kameenui, 2001). In addition, a series of risk indicators will be created for each screening measure that can be used to predict student failure on the high-stakes reading assessment through the use of a large, urban sample.

The following research questions were addressed in this study:

1. Can the DORF and DRA measures accurately predict student performance on the statewide reading assessments?
2. What scores (i.e., risk indicators) on both the DORF the DRA are predictive of student failure on the statewide reading assessment?

CHAPTER II

Literature Review

The previous chapter identified the great importance reading has on an individual's functioning in society in the United States today. Also discussed were the research initiatives and subsequent program implementation by the United States government meant to improve reading outcomes for American children. The monitoring of the effectiveness of these programs by mandatory statewide assessments of reading performance was also discussed. Finally, Chapter I described the use of R-CBM techniques as a way of providing necessary information to school districts regarding student performance in reading prior to administration of the statewide assessments. R-CBM procedures can provide screening-type information that can be used to predict student performance on statewide assessments and, in turn, alter student programming by providing supplemental reading support programs or intervention services when needed.

Chapter II will discuss additional information relating to the development and usefulness of R-CBM in schools and the importance of screening measures as tools for predicting student performance on an outcome measure and providing information school personnel can use to make educational decisions. Specifically, this chapter will focus on two screening measures, DORF and the DRA, and how they are used to predict student performance on outcome measures. Furthermore, the use of high-stakes state reading assessments and how the data obtained from these assessments relates to the data obtained from the screening measures will be addressed. Specifically, the chapter will address how screening data can be used to predict student performance on high stakes state reading assessments. Finally, the importance of benchmarks and how they are developed, used, and assessed will be discussed.

Curriculum-Based Measurement of Reading

Curriculum-based measurement was originally developed by Stanley Deno and colleagues at the University of Minnesota the in the early 1980s as a method for measuring student growth in a variety of academic skills. The original purpose of Deno's research was to develop a system of measurement that could be used by special education teachers to make accurate decisions as to when and how to modify a student's instructional programming (Deno, 1985). CBM would, therefore, provide teachers with a tool that would enable them to frequently monitor student academic progress so that those instructional changes could be made (Deno). When developing the measures, it was deemed important that they meet four established criteria in order to be considered effective. The measures needed to be: (a) reliable and valid; (b) simple and efficient; (c) easily understood by teachers, parents, and students; and (d) inexpensive to enable the use of multiple forms (Deno).

CBM can be used to assess academic performance in several different academic areas including reading, mathematics, written expression, and spelling (Marston, 1989). The goal of curriculum-based measures of reading (R-CBM) is to accurately measure student reading performance. This goal can be accomplished by measuring the fluency and accuracy of a student's oral reading of a short passage of text. The number of words read correctly in one minute (WRC) is calculated for each student. Reading aloud from text has been demonstrated as a reliable and valid measure of reading ability that can be used to monitor student growth in reading throughout the elementary years (e.g., Deno, 1985; Deno, 2003; Fuchs & Fuchs, 1999; Fuchs, Fuchs, Hosp, & Jenkins, 2001). Reading aloud from text was also demonstrated to be a valid way to discriminate between students enrolled in special education programming and those not enrolled in special education programming (Fuchs & Deno, 1981). Research also supported

the use of R-CBM in special education programming decisions, screening, establishment of student goals, progress monitoring, and to inform instructional changes (Deno, 1985).

The technical adequacy of R-CBM has been strongly supported through a series of studies (Marston, 1989). R-CBM was found to correlate strongly with other commonly used norm-referenced tests of reading achievement with Pearson correlation coefficients ranging from .73 to .91, with most coefficients above .80. Also, correlations between R-CBM and oral reading performance on basal reader mastery tests were found to be .84. R-CBM correlated highly with teachers' judgments of reading performance. The median correlation between these two measures was .86. Test-retest reliability estimates across several studies ranged from .82 to .97 with most exceeding .90. Alternate form estimates ranged from .84 to .96 with most correlations above .90. Inter-rater agreement coefficients were very high at .99 (Marston, 1989). Taken together, the studies reviewed provide strong support for the technical adequacy of R-CBM. Concerns regarding R-CBM's utility centered around issues relating to its low face validity (i.e., measures do not formally assess the ability of the student to understand the passage (Fuchs, Fuchs, Hosp, & Jenkins, 2001). Additional concerns regarding possible cultural or gender biases in R-CBM were also noted (Kranzler, Miller, & Jordan, 1999).

In order to measure the overall goal of reading, which most consider to be comprehension of text, R-CBM measures should be related to growth in text comprehension. Several studies were conducted to investigate the validity and reliability of R-CBM as an indicator of reading outcomes, such as comprehension (e.g., Fuchs, Fuchs, & Maxwell, 1988; Shinn, Good, Knutson, Tilly, & Collins, 1992). Fuchs, Fuchs, and Maxwell (1988) performed a study to investigate the use of a series of informal reading measures as indicators of reading comprehension. The study included 70 boys who ranged in age from 9 to 15 years. The participants were identified as

students with learning disabilities, emotional disturbances, or mental retardation. The students were administered four informal reading measures including a comprehension question test, a passage recall measure, an oral reading test, and a cloze procedure. They were also administered two reading subtests from a global norm-referenced achievement test. One subtest assessed phonetic and structural analysis with consonants and vowels while the other assessed comprehension of text. Performance on the informal reading measures was then compared to performance on the subtests of the norm-referenced achievement test.

Results indicated the correlation between the oral reading test and the norm-referenced test of text comprehension was significantly higher than the correlation between each of the other three informal reading measures and the reading comprehension subtest. Thus, this study supported the use of oral reading rate as a useful method for monitoring reading growth and reading comprehension. However, due to the low face validity of oral reading measures, results suggested that the remaining three informal reading measures were adequate indicators of reading comprehension that could be utilized if practitioners were uncomfortable using oral reading measures.

Similarly, Shinn, Good, Knutson, Tilly, and Collins (1992) performed a confirmatory factor analysis with third (N = 114) and fifth (N = 124) grade students in which they investigated the relationship between R-CBM and reading comprehension. The study investigated the theoretical role of fluency in reading by comparing four pre-established models of reading including (a) a unitary model where decoding, fluency, and reading comprehension were not distinct components of reading; (b) a two-factor model involving decoding and reading comprehension, where fluency was considered a component of decoding; (c) a second two-factor model involving decoding and comprehension, where fluency was considered a component of

comprehension; and (d) a three-factor model where decoding, comprehension, and fluency were considered separate constructs.

The researchers administered a series of measures to groups of third and fifth grade students. These measures, meant to assess different aspects of reading, were (a) two R-CBM passages taken from the district's most frequently used textbook, (b) a list of phonetically regular words and phonetically regular nonsense words that students were asked to read aloud, (c) a written retell task based on a 400 word folktale, (d) a cloze task based on a 400 word folktale, and (e) the Reading Comprehension subtests from the Stanford Diagnostic Reading Test (SDRT). Results indicated the three-factor model was supported for both third and fifth grades; however, this model did not explain the relationship between fluency and the other reading constructs most simply. For third grade students, the unitary model could not be rejected. The fluency measures had higher factor loadings in the single-factor model than the factor loadings for the more conventional reading comprehension measures. For example, factor loadings in the single-factor model were .68 for the written retell task and .90 for the oral reading fluency task. For fifth grade students, the two-factor model of reading where fluency represented decoding was supported. Fluency measures were also found to correlate as high or higher with the reading comprehension construct as the measures meant to assess reading comprehension in the study. However, all measures of reading comprehension included in this study contained a written component thus creating a potential confound in the data and limiting the applicability of these results. Despite this potential confound, the study supported the inclusion of fluency in theoretical models of reading and supported the ability of fluency measures to assess both lower level and higher level reading skills, including comprehension (Shinn et al.).

The issue of possible cultural or gender bias in R-CBM was addressed by Kranzler, Miller, and Jordan (1999) who conducted a study to investigate the properties of R-CBM across a variety of racial, ethnic, and gender groups. Results indicated potential bias for racial and ethnic groups in grades four and five and gender groups for grade five. No bias was indicated for grades two and three. Specifically, R-CBM tended to overestimate the reading comprehension of African American students and underestimate the reading comprehension of Caucasian students. Furthermore, results for grade five indicated R-CBM performance overestimated the reading comprehension of girls and underestimated the reading comprehension of boys. Thus, questions regarding the usefulness of R-CBM as a screening tool were raised (Kranzler, Miller, & Jordan, 1999).

Hintze and colleagues (2002) replicated and extended the work of Kranzler and colleagues (1999). The predictive bias of R-CBM with African American and Caucasian children in grades two through five was investigated. Results from a series of hierarchical multiple regression analyses indicated no bias such that no overestimation or underestimation of performance based on the R-CBM measures was noted. Results of this study contradict those of the Kranzler et al. study and support the use of R-CBM as a valid tool for predicting overall reading performance in both African American and Caucasian elementary students (Hintze, Callahan, Matthews, Williams, & Tobin, 2002). Given the mixed evidence provided by these studies, firm conclusions regarding the use of R-CBM with different ethnic and gender groups cannot be drawn.

Screening

A screening system in the primary grades must accomplish three goals. First, it should be able to measure and account for growth in a variety of skills related to early literacy. Second, it

should be able to predict student success or failure on an outcome measure, such as a high-stakes assessment. Finally, the screening system should be able to provide an instructional goal (i.e., benchmark) that, if met, will be highly indicative of future reading success. (Good, Simmons, & Kameenui, 2001). R-CBM is an example of a screening measure that satisfies these three criteria.

The type of screening measure used can play a vital role in accurately determining which students will require intervention services. Essentially, four possible outcomes can result from an assessment with a diagnostic screening measure (Davis, Lindo, & Compton, 2007). First, the screening measure identifies a child as "at-risk" for reading failure when that child is actually at risk (i.e., he or she will require intervention support in order to succeed in reading). This outcome is known as a "true positive" (TP) such that the outcome obtained from the screening measure is commensurate with the reality of the situation (i.e., the child does need additional support). The second possible outcome from the screening measure is known as a "true negative" (TN). In this case, a child is determined not to be "at-risk" for reading failure and this decision is commensurate with the child's true abilities (i.e., he or she will not need additional academic intervention support in order to be successful in reading).

Two incorrect outcomes also can be obtained from screening measures. The first of these is known as a "false positive" (FP). A false positive occurs when a child is determined to be "at-risk" for reading failure by the screening measure; however, he or she would have been able to be successful in reading without intervention support (i.e., the child is not actually "at-risk" for failure). False positives inflate the number of students identified as needing intervention services and thus stress the school's resources unnecessarily because these children would be able to be successful without the additional support. The final possible outcome is known as a "false negative" (FN). These students are determined not to be "at-risk" by the screening measure but

later go on to experience difficulty in reading. Because these students were not accurately identified by the screening measure, they do not receive the early intervention services required for reading success. The case of a "false negative" should be minimized due to the severely negative results of this incorrect screening (i.e., a child who needs support does not receive it).

The usefulness of a diagnostic screening measure hinges on its ability to accurately yield a high percentage of "true positives" and "true negatives" while at the same time limiting the number of incorrectly identified "false negatives" and "false positives." Schools attempt to accomplish this goal by establishing "cut-scores," which are screening measure scores that differentiate students into categories of "at-risk" or "not at-risk." These scores can be established and adjusted based on the school's particular population and the school's resources. Overall, the school should attempt to establish cutoffs that provide a balance between true positives and false positives by weighing the cost and benefit of incorrectly identifying students as needing or not needing intervention services.

In a prevention-oriented system, it is important that information gained from screening systems be used to inform instructional decisions. In doing so, it is natural that instructional changes may follow. Interventions provided to students in need may, in turn, alter the predictive validity of the screening measure. Thus, continued screening periods to assess the performance of each student are necessary to ensure that interventions put in place are causing the desired change toward increased reading ability in students.

Screening methods used may differ based on each school district's specific needs and resources. Two commonly used systems, DIBELS and the DRA, are designed to accomplish the aforementioned goals. The two systems efficiently provide information related to early literacy skill development to district personnel. Both measures can be used multiple times throughout the

school year to monitor student progress toward academic outcome goals. Using instructional goals or benchmarks provided by these screening systems, student progress data can be utilized to predict the performance of individual students on high-stakes reading assessments. This information can then be used to inform instructional changes in order to ensure students continue to progress toward success in reading.

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Oral Reading Fluency

Developed by a team of researchers at the University of Oregon, the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) is a comprehensive system for curriculum-based assessment in reading that includes a series measures meant to provide quick, reliable, and valid measurements of the early skills that students need to master in order to be successful readers (Good & Kaminski, 1996). When these early skills have been mastered, an individual will be able to read a passage fluently and understand its meaning. This final culminating task as measured by DIBELS is known as Oral Reading Fluency (DORF). DORF is a standardized measure meant to assess an individual's accuracy and fluency when reading a short written passage. (Good, Kaminski, Simmons, & Kameenui, 2001, p.10).

The Developmental Reading Assessment

The Developmental Reading Assessment (DRA) is another screening measure used to evaluate student performance in reading. The DRA was first developed by the Upper Arlington School District in 1986 in response to a document published by the United States Department of Education entitled *A Nation at Risk* (National Commission on Excellence in Education, 1983). The pilot version of the DRA was completed in 1988 and began to be used as an assessment tool to identify students at risk for reading failure in grades kindergarten through three in school districts in Ohio (Beaver, 2004). The DRA was revised several times and an extended version

was created in 2000 which can be used to assess reading skills in students in grades four through eight.

The DRA measures three different components of reading including engagement, fluency and accuracy, and comprehension. It is designed to be both administered and interpreted by teachers. It uses authentic texts to measure student performance and can be administered on an annual to semiannual basis or more frequently with struggling readers in order to monitor growth (Beaver, 2004). The DRA K-3 includes 20 levels of text difficulty that range from level A (easiest) to level 40 (most difficult). During the assessment, the teacher notes the student's oral reading ability and responses to comprehension questions about the presented text (Beaver, 2004). An oral fluency rate is not calculated for students below grade four.

The DRA differs from DORF in several ways, including that it does not provide a measure for oral reading fluency. However, the DRA focuses on comprehension by utilizing questions designed to assess how much information the student was able to glean from the text (Beaver, 2004). There has not been much research on the DRA since its creation; therefore, other uses of the DRA such as its use as a predictive tool for student performance on high-stakes tests of reading achievement are not fully known. Despite the lack of evidence to its effectiveness as a tool for monitoring reading achievement, the DRA is used in over 30,000 school districts across the United States for this purpose (Pearson Learning Group, 2003).

High Stakes Assessment

Following the implementation of federal initiatives related to student performance in reading (e.g., No Child Left Behind, Reading First), there has been an increased demand not only for the use of evidence-based reading interventions, but also for adequate assessment of student reading levels and accountability for school districts. In order to meet these demands, most states

have developed some type of comprehensive reading examination that is administered annually to students (McGlinchey & Hixson, 2004).

New York State is one of the many states that has altered its curriculum based on the recent changes in federal legislation. After the NCLB legislation was enacted, New York State revised its Language Arts Core Curriculum, a document that specifies what New York State students need to learn in reading and language arts (New York State Department of Education, 2005). The Language Arts Core Curriculum includes four standards students must meet. These standards delineate that students should be able to read, write, listen, and speak for information and understanding, literary response and expression, critical analysis and evaluation, and social interaction (New York State Department of Education, 2005). Attainment of these standards is assessed with the New York State English and Language Arts Examination on a yearly basis (New York State Department of Education, 2005).

Using Screening Measures to Predict Performance on High-Stakes Assessments

With the increased focus on student performance on high-stakes tests of achievement throughout the past several years, many studies have been conducted to assess the relationship between curriculum-based measures, such as ORF, and statewide tests of reading achievement (e.g., Good, Simmons, & Kameenui, 2001; McGlinchey & Hixson, 2004). Research on the topic has focused on prediction of either passing or failing the statewide assessment, with most studies aimed at predicting a passing score. In this case, researchers have developed benchmarks, or cut-off scores that, if attained, indicate the child is likely to achieve a passing score on the outcome measure. In the case where a study is aimed at predicting student failure on an outcome measure, risk indicators are developed, or scores that are indicative of failing the outcome measure if the student scores at or below that designated score. Thus, as opposed to monitoring changes in

student performance (as is done in progress monitoring), utilizing R-CBM for screening purposes provides information that is useful in predicting later student performance on an outcome measure and informing instructional modifications to assist students in attaining designated outcome goals.

In 2001, Good and colleagues (2001) tested established benchmarks for both DIBELS (early literacy measures) and DORF in an urban district in the Northwest United States (Good, et al.). This longitudinal study utilized the Oregon Statewide Assessment (OSA) as the high-stakes measure of student reading achievement. Participants were four cohorts of elementary-aged students from six schools. Five of the six elementary schools in the study were eligible for Title I services, 10 percent of the students were from a minority group, and 37 to 63 percent of the students received free or reduced-cost lunch.

Benchmarks based on a trajectory of desired progress toward an outcome measure were used. The initial benchmark of 40 WRC by spring of first grade was used to identify benchmarks for second and third grades. Benchmarks for spring of second and third grades were determined to be 90 WRC and 110 WRC, respectively. Students were administered three different ORF passages in the spring of their first, second, and third grade years. The median scores on the three passages administered for each of the three years were compared to the students' performances on the OSA. The benchmarks were then tested to determine their appropriateness in predicting students who were likely to succeed on the OSA.

Results indicated 96 percent of the students who reached the spring benchmark for third grade (110 WRC) met or exceeded the expectations of the OSA (Good, et al., 2001). Conversely, only 28 percent of students who did not attain the ORF benchmark by the spring of third grade were able to meet or exceed expectations on the OSA. Spring ORF performance for the cohort of

students going from second to third grade was not available. Thus, the second-to-third grade linkage was not examined (Good et al., 2001).

Crawford and colleagues attempted to establish a predictive link between R-CBM and the OSA using chi square statistics (Crawford, Tindal, & Stieber, 2001). Students were administered R-CBM passages derived from the district's basal reading series. ORF scores were then correlated with scores on the OSA to determine which scores for ORF best predicted later performance on the OSA. Previously established norms based on the work of Hasbrouck and Tindal (1992) were used to classify students into groups based on their ORF and OSA scores. Results indicated a direct relationship between ORF scores and performance on the OSA. For third grade students, 119 WRC was determined to be the ORF score needed to predict passing on the statewide reading test with 94 percent of students scoring 119 WRC or higher later going on to pass the OSA. For second grade students, a score of 72 WRC on the ORF measure resulted in a 100 percent passing rate on the OSA, which was taken during third grade. However, the small sample size of this study (N = 51) may have led to less accurate benchmarks that differ from similar studies, particularly for second grade. Overall, this early study also provides support for the use of ORF measures as predictors of later performance on high-stakes tests of reading achievement (Crawford, et al.).

Sibley and colleagues (2001) replicated and extended the Good at al. (2001) study by investigating the utility of the benchmarks established in the Good at al. study for a suburban school district in Illinois. Students were administered ORF probes twice per year. Student performance on the probes was then correlated with performance on the Illinois Standards Achievement Test (ISAT). Growth rate analysis based on slope data developed by Fuchs et al. (1993) was used to evaluate the appropriateness of the benchmarks. Results indicated support for

utilizing established benchmarks as predictors of student performance on the ISAT. An ORF score of 90 WRC and 110 WRC was supported for second and third grade spring ORF benchmarks respectively. Thus, this study provides further support for the ability of student ORF scores to predict later performance on high-stakes tests of reading achievement and it provides support for the particular previously-established benchmarks for second and third grade students (Sibley, Biwer, & Hesch, 2001).

Similarly, Stage and Jacobsen (2001) conducted a study to determine whether student performance on ORF probes would predict later performance on the Washington Assessment of Student Learning (WASL), a statewide test of reading achievement administered to fourth grade students. The WASL is composed of multiple choice, short-answer, and extended response questions. The researchers used hierarchical linear modeling (HLM) growth curve analysis to investigate the relationship between changes in individual students' ORF performance over time (i.e., slope) and the WASL. In order to determine the number of words read correct at each interval period, the students' slopes were converted back into words read correct per minute and an analysis of variance was used to determine the cut-scores based on the 95 percent confidence interval for the number of words read correct per minute and WASL level performance.

Results indicated scores on ORF measures obtained as early as September of the testing year could accurately predict those students who were "at risk" for failing the WASL, which was administered in May of that year (Stage & Jacobsen, 2001). In addition, the HLM growth curve analysis indicated ORF level scores were more accurate predictors of performance on the WASL than the growth in a student's ORF abilities over the year (i.e., slope). The authors also noted that the ability to predict failure on the WASL was increased by 30 percent when ORF cut-scores were used resulting in 74 percent of students being correctly classified as "at risk" or "not at

risk" for failing the WASL. (Stage & Jacobsen, 2001). Benchmarks predicting passing for fourth grade students on the WASL were 107 WRC for the fall benchmarking period, 122 WRC for the winter benchmarking period, and 137 WRC for the spring benchmarking period. The results of this study provide support for the use of R-CBM by school districts to aid in early identification and intervention for students who are less likely to succeed on state reading tests. In addition, these results provide specific scores that are able to accurately identify students at risk for failing the WASL (McGlinchey & Hixson, 2004).

Limitations of the Stage and Jacobsen (2001) study include a lack of cultural and socioeconomic diversity in the sample. Ninety percent of the participants were of European American descent and only fifteen percent of the participants were eligible for free or reduced-cost lunch, an indicator of socioeconomic status. Furthermore, only fourth grade students were assessed. Therefore, the generalizability of these results to more culturally and economically diverse school districts as well as to other grade levels is limited. Finally, since the WASL is only administered in Washington State, these results are somewhat limited in their application such that generalizations cannot be made to similar assessments given in other states.

McGlinchey and Hixson (2004) replicated and extended the results of Stage and Jacobsen (2001) by assessing the predictive power of ORF measures on the Michigan Educational Assessment Program (MEAP). This multiple-year study involved approximately 11,000 students assessed over eight years. Fifty-two percent of participants were non-Caucasian and 60 percent qualified for free or reduced-cost lunch. Assessment with the R-CBM probes took place in grade four, one month prior to administration of the MEAP. The MEAP was administered in October of fourth grade for the first three years of the study and February of fourth grade for the remaining four years of the study (McGlinchey & Hixson., 2004).

An ORF score of 100 WRC was identified as the score that most accurately differentiated student performance on the MEAP such that students scoring at or below this value were likely to fail the MEAP, while students scoring above 100 WRC were likely to pass. This score correctly classified 74 percent of students into categories of "likely to pass" and "likely to fail." Therefore, a moderate relationship existed between student performance on the ORF probes and performance on the MEAP. Thus, this study provides support for the link between ORF and high stakes assessments of reading achievement, particularly through its use of a more culturally and socioeconomically diverse sample of students that were assessed longitudinally (McGlinchey & Hixson, 2004).

Comparison of the results of previous studies suggests consistency among the developed benchmarks. Since the WASL and MEAP were administered at different times of the school year, comparing the benchmark values for assessment periods immediately prior to assessment with the state reading tests indicates similar benchmarks for both assessments (107 WRC for the WASL and 100 WRC for the MEAP). The McGlinchey and Hixson (2004) study also alleviated some of the geographic generalizability issues associated with the Stage and Jacobsen (2001) study by utilizing a different state reading assessment. However, the use of only one grade level limits the generalizability of these results to other grades. Furthermore, variability among testing conditions on the MEAP and testing modifications made for special education students were not fully known.

In order to investigate the most effective statistical method for determining cut scores, Silberglitt and Hintze (2005) conducted a study involving over 2,000 students from a rural/suburban district in Minnesota. Student performance on ORF probes was compared to performance on the Minnesota Comprehensive Assessment (MCA), a statewide test of

achievement administered to all students in the spring of third grade. Four data analysis methods were used to evaluate and define cut-scores in this longitudinal study. Discriminant analysis, which determines the probability of membership in a group by examining a set of variables that describe a population, was used to group those who did and did not pass the MCA based on ORF scores. The equipercentile method applied the percentage of students scoring below a passing score on the MCA to those students' ORF scores to arrive at an equivalent percentile score on ORF. Logistic regression, which determines the likelihood of membership in a certain category, used each student's MCA score as the dependent variable and the ORF score as the independent variable. Finally, Receiver Operating Characteristic (ROC) curve analysis, in which the sensitivity and specificity of a predictor variable is plotted for all possible values of the cut score, resulted in the creation of a graph that analyzed the sensitivity and specificity to determine the strength of the predictor. This ROC curve can also be used to determine the diagnostic accuracy of the cut scores.

Results of this study suggested logistic regression and ROC curve analysis were the most effective methods for evaluating and defining cut scores. Although the diagnostic accuracy of cut scores generated by ROC curve analysis was not as high as with linear regression, ROC curve analysis yielded higher negative predictive power and its flexibility provided strong diagnostic accuracy thereby generating results similar to those produced by linear regression. Thus, ROC curve analysis was determined to be most useful way of evaluating and defining cut scores.

A study by Hintze and Silberglitt (2005) further extended the results of the previously discussed studies by using ROC curve analysis to create benchmarks that would accurately predict student success on a state test of reading achievement. R-CBM data on 1,766 elementary students from a district in the northern central region of the United States was collected over

three years. Each student was assessed a total of eight times throughout the three-year period. Statistical analyses then compared R-CBM cut-scores with student performance on the reading portion of the MCA (Hintze & Silberglitt, 2005).

The authors used ROC curve analysis to develop benchmarks for the MCA. Benchmarks were developed for first, second, and third grade students that could accurately predict the likelihood of passing the MCA in third grade. The benchmarks for second grade students were 41 WRC in fall, 71 WRC in winter, and 88 WRC in spring. The benchmarks for third grade students were: 68 WRC in fall, 93 WRC in winter, and 109 WRC in spring. This study provides support for the use of R-CBM in predicting student success on statewide standardized tests of reading achievement. Specifically, this study further supports the use of ROC curve analysis as an effective method for analyzing ORF data and developing appropriate benchmarks for predicting student performance on high-stakes assessments of reading achievement.

The benchmarks identified in the Hintze and Silberglitt (2005) study coincide with those identified in previously discussed studies. The benchmark for spring of third grade was designated as 109 WRC. This benchmark is similar to the benchmark identified in both the Good et al. (2001) study (110 WRC), which predicted success on the OSA as well as the Stage and Jacobsen (2001) study (107 WRC), which predicted success for fourth graders on the WASL. The Hintze and Silberglitt (2005) study also provides further support for the consistency of these results across a variety of state tests of reading achievement. Furthermore, this study included a more economically diverse population than previous studies indicating further generalizability across differing school districts.

Summary

Given the numerous initiatives aimed at increasing early literacy development in elementary school children, it is clear that development of effective methods for assessing and monitoring this progress is necessary. Accountability of school districts, as assessed by high-stakes statewide achievement tests, provides districts with the incentive to implement screening systems that can not only provide information relating to the prediction of student performance on outcome measures, but also provide data that can inform instructional decisions and identify students in need of increased academic support. Several studies have addressed the issue of creating benchmarks or risk indicators based on R-CBM procedures that are effective at predicting student performance on high-stakes tests of reading achievement. The present study will replicate and extend these previous studies in order to provide more information relating to the use of risk indicators to predict later student performance on a statewide reading assessment.

CHAPTER III

Method

Participants

Participants in the study included 195 second grade students enrolled in four different elementary schools in a midsize urban school district in the Northeastern United States. The study involved students from 22 different second grade classrooms. The four elementary schools within the district were all involved in the Reading First Program. Students in the current study were enrolled in second grade during the 2004-2005 school year, participated in the DORF and DRA assessments during second grade, and took the NYS ELA examination during the following school year.

The district included in the study is composed of approximately 36,500 students enrolled in 57 schools throughout the district. Seventy-eight percent of students within the district receive subsidized meals. Eight percent are considered English Language Learners, and 15 percent receive special education services. The racial and ethnic makeup of the district is 65 percent African American, 20 percent Hispanic, 14 percent Caucasian, and 2 percent Native American, Asian, or another race or ethnicity.

Measures

Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency Probes (DORF)

DORF is an assessment tool that is a form of R-CBM designed to measure an individual's ability to accurately and fluently read a short passage. Performance is measured by recording the number of words read correctly per minute (WRC) as the student reads aloud to the examiner. The examiner marks the number of words read incorrectly or those words the student does not read correctly within three seconds. Words the child self-corrects within three seconds

are marked correct. The examiner then calculates the number of words read correctly from the passage by subtracting the number of incorrect words from the total number of words read in one minute. Three probes were administered to each child at each assessment period (fall, winter, and spring). The median number of WRC and the median number of errors were then calculated and recorded for each student during each assessment period.

Many studies over the past 25 years have addressed the technical adequacy of R-CBM measures. Some of the earlier studies on this topic are summarized in a review of the literature conducted by Marston (1989). Validity of R-CBM measures is supported by high correlations among ORF measures and commonly used criterion tests of reading. Deno and colleagues (1982) found oral reading fluency to be a valid measure of reading ability. Correlations among oral reading fluency measures and criterion tests of reading ranged from .73 to .91, with most coefficients exceeding .80 (Deno, Mirkin, & Berttram, 1982). Other studies involving additional published measures of reading skills reported correlations ranging from .63 to .90, with most coefficients exceeding .80. Fuchs and Deno (1981) reported median correlations between ORF measures and teacher judgement of student reading progress to be .86 (Fuchs & Deno, 1981).

In regard to reliability of R-CBM measures, studies summarized by Marston (1989) indicated test-retest reliability coefficients ranging from .82 to .97 with most coefficients exceeding .90. Reliability coefficients for parallel forms ranged from .84 to .96, with most exceeding .90. Interrater reliability coefficients were reported to be .99. Overall, the data accumulated over many years of research provides strong support for the reliability and validity of R-CBM measures.

Benchmark scores for DORF are provided through the assessment materials. These scores are used to compare a student to others in the same grade and are an established standard

of performance that can be used to indicate that student's likelihood of reading success. Benchmarks for each assessment period for second grade students on the DORF are as follows: fall benchmark = 44 WRC, winter benchmark = 68 WRC, and spring benchmark = 90 WRC.

Developmental Reading Assessment (DRA)

The DRA is a teacher-administered assessment designed to measure literacy skills for students in grades kindergarten through eight (Pearson Learning Group, 2003). Administration time for the DRA is approximately 10 to 20 minutes depending on text difficulty and the appropriateness of the difficulty level of the selected text to the student's independent reading level. Teachers complete Observation Guides in order to evaluate student reading performance while students read from short texts ranging in difficulty from level A (easiest) to level 44 (most difficult).

The teacher selects the text that is believed to be closest to the child's reading level. The teacher shows the child the text and asks him or her to make a prediction about the story based on either the pictures (for levels 3 through 16) or on information obtained by reading the first several paragraphs aloud (for levels 18 through 44). Students reading above level 2 are then asked to retell the story while the teacher uses scripted questions to assess comprehension of the text. Information collected on the Observation Guide, is then used to determine the student's Independent Reading Level.

Information relating to the technical adequacy of the DRA is provided in the technical manual. Two forms of reliability, test-retest and scoring, have been investigated for the DRA. Weber (2000) investigated the test-retest reliability of the DRA following the administration of the DRA to 306 students by 68 first through third grade teachers. Students were assessed with the DRA twice during a three week period. Results of both test administrations were correlated

indicating test-retest reliability coefficients between .92 and .99 for the first and second administrations of the DRA (Weber).

Williams (1999) investigated the scoring reliability of the DRA by examining the inter-rater agreement of 87 teachers from 10 different states. Each teacher assessed at least three different students from his or her class and audio taped the assessment session. The original teacher and two blind assessors then scored each audio tape. Correlational analyses indicated inter-rater agreement between the original teacher and the first rater to be .80, which is considered barely adequate for screening measures. Inter-rater agreement was even lower for all three raters (.74) (Williams, 1999).

Interscorer agreement, or the ability to ensure that a student's score would be constant if rated by any teacher on any given day, was investigated by Weber (2000). Ten teachers observed an expert administer the DRA to four different students. Each teacher scored the students' accuracy with oral reading. Percents of agreement with the expert (within 2%) were high for most assessment levels indicating high observer validity. Assessment levels A through 3 demonstrated 100 percent interscorer agreement. Levels 4 and 6 demonstrated 90 percent agreement, and levels 18, 24, 28, 40, and 44 demonstrated 100 percent agreement. Only level 8 demonstrated lower interscorer agreement at 70 percent (within 2%). However, when asked to score the students' comprehension, the raters percents of agreement with the expert were much lower with interscorer agreement within one score point with the expert ranging from 14.3 percent to 40 percent. (Weber).

In the present study, the district involved utilizes its own benchmarks to which individual student's scores on the DRA can be compared. Benchmark scores for the DRA that are indicative of an increased likelihood of reading success are as follows: fall benchmark = level 18 and spring

benchmark = level 28. Benchmarks for the district are based on assessment at the "instructional" level, meaning students were assessed with texts that were more difficult than what the child would be expected to read on his or her own, but would be appropriate for classroom instruction. In contrast, assessment at a reading level at which the child was successful reading on his or her own would be assessment at the "independent" level.

New York State English Language Arts Examination (NYS ELA)

The 2006 NYS ELA was administered on two consecutive days from January 9, 2006 through January 13, 2006. For students in grade three, the test is made up of 24 multiple choice questions and 4 constructed response questions based on information contained in short passages. The constructed response items require the students to formulate written responses to questions based on the passages. Items contained in the NYS ELA are designed to measure the skills, concepts, and processes taught in New York State schools. Teachers provide standardized instructions read aloud. Students are instructed to read or listen to the passages and answer the corresponding questions. Students indicate their answers by filling in circles on an answer sheet. Third grade students have 40 minutes to complete the reading section (Day 1) and 35 minutes to complete the listening section (Day 2).

Scoring takes place at a designated site by qualified teachers and administrators. The scoring of the constructed response items was based on the scoring guides developed by CTB/McGraw-Hill Handscoring. Development of these scoring guides included input from the New York State Department of Education and New York State teachers. Student responses were discussed and reviewed and a consensus score was agreed upon. Test booklets were randomly dispersed through scoring sites so as to avoid any bias in test scoring. Students earn performance level score ranging from 1 to 4 where students who score within levels 1 and 2 are considered

not to be meeting grade-level expectations in reading and students scoring within the level 3 or 4 range are considered to be meeting grade-level expectations in reading (New York State ELA Technical Report, 2006).

Content validity of the NYS ELA is carefully matched to specific standards in the curriculum. NYS teachers are involved in the development of the test and reviewed the field tests to assess the degree to which test items align with curriculum standards. Construct validity is also supported for the ELA with reliability coefficients ranging from .82 to .89. Finally, high internal consistency has been evidenced, with a Chronbach's alpha of .85 (New York State ELA Technical Report, 2006).

Procedures

The archival data set was obtained from a staff member of the school district. The data set contained student scores on both screening measures (DORF and DRA) as well as each student's score on the NYS ELA administered in 2006. DORF measures were administered to all participants in the fall, winter, and spring of second grade by trained teachers and other faculty members from the district. The exact training procedures and methods for ensuring reliability of the DORF data collection are not fully known because the data set was archival. The DRA was administered to each participant once during the fall and twice during the spring of second grade by trained teachers and faculty members from the district. Data from the first spring administration of the DRA were excluded from the study and data from the second administration were used because more students were present for the DRA assessments during the second spring administration period. Again, exact training procedures and methods for ensuring reliability of the DRA data collection are not fully known because the data set was

archival. The NYS ELA was administered in January of third grade. Administration instructions were provided by the New York State Department of Education.

Confidentiality

The data analyzed in the current study was a portion of an archival data set collected by staff members from the school district. In order to maintain confidentiality, student names were removed from the data set prior to analysis by the researchers. Furthermore, confidentiality agreements prepared by the school district were signed by the researchers to ensure confidentiality of the database and information therein.

Data Analysis

Descriptive statistics were calculated for the data set and correlational analyses were conducted. Receiver Operating Characteristic (ROC) curves were created to assess the diagnostic accuracy of the screening measures over a variety of possible cut-cores (Streiner & Cairney, 2007). The ROC curves were created by plotting the sensitivity (i.e., the screener's ability to identify students who were truly "at-risk" for not passing the ELA) against 1-specificity (i.e., 1 – the screener's ability to identify students who were truly not "at-risk" for not passing the ELA) across a range of possible cut-scores. These ROC curves were then used to determine the accuracy of each screening measure in predicting later student performance on the ELA. Sensitivity and specificity values generated by the statistical software were also used to calculate the cut-scores that were deemed most effective at predicting later student performance on the ELA.

In order to determine the most appropriate cut-score for each administration period of the two screening measures, values for Positive Predictive Power (PPP = (base rate X sensitivity) / (((base rate X sensitivity) + ((1 – base rate) X (1 – specificity))), Negative Predictive Power

(NPP = ((1 – base rate) X specificity) / (((1 – baserate) X sensitivity)) + (base rate X (1 – sensitivity))), and Correct Classification (CC) were calculated (Glover & Albers, 2007). The positive predictive power is estimated by first calculating the product of the base rate and sensitivity. This value is then divided by the sum of the product of the base rate and the sensitivity and the product of one minus the base rate and one minus the specificity. The negative predictive power is estimated by first calculating the product of one minus the base rate and the specificity. This value is then divided by the sum of the product of one minus the base rate and the sensitivity and the product of the base rate and 1 minus the sensitivity.

The correct classification (CC) index rating was calculated for each cut-score by adding the total number of students who were identified "at-risk" for failing the ELA at that cut-score and were not successful on the ELA at that score (true positives) with the number of students who were not identified "at-risk" for failing the ELA at that cut-score and did go on to pass (true negatives) and dividing that value by the total number of students administered the screening measure at that assessment period.

Identifying appropriate cut-scores for each administration period for both screeners involved choosing the value that provided the best compromise between sensitivity and specificity. The CC value was used as an additional source of information to determine which cut-score was most appropriate at each assessment period for both screeners. For each assessment period of both DORF (fall, winter, and spring) and the DRA (fall and spring), the most appropriate cut-score was derived based on the sensitivity and specificity data. Comparisons between these cut-scores and the established benchmarks for the DORF and the district benchmarks for the DRA were then made.

CHAPTER IV

Results

Descriptive Statistics

Table 1 contains the descriptive statistics for all children on the three administrations of DORF, the two administrations of the DRA, and the ELA performance level scores. The distributions for each assessment were examined to determine normality. The distribution for the fall administration of DORF was slightly positively skewed with more children scoring in the lower range of number of words read correct per minute. The winter and spring distributions for DORF were more normally distributed with the majority of children falling within the average range of numbers of words read correct per minute on the DORF probes. Examination of the distribution for the fall administration of the DRA suggested a normal distribution of scores. The distribution of scores for the spring administration of the DRA appeared negatively skewed suggesting more consistency among the scores of the children and a smaller range of performance for this assessment period.

Table 1

Descriptive Statistics for Total Sample (N = 195)

Variable	N	Mean	(SE)	Min.	Max.	Skew.	SE Skew.
ELA SS	195	651.66	2.34	587	780	.818	.174
ELA PL	195	2.35	.05	1	4	-.189	.174
ORF F	183	46.26	1.77	0	151	1.133	.180
ORF W	184	71.28	2.21	10	186	.766	.179
ORF S	189	85.02	2.28	25	211	.721	.177
DRA F	169	10.38	.24	1	18	-.208	.187
DRA S	183	14.85	.22	1	20	-.648	.180

Note. ELA SS = English Language Arts Examination Standard Score; ELA PL = English Language Arts Examination Performance Level; ORF F = Oral Reading Fluency Fall; ORF W = Oral Reading Fluency Winter; ORF S = Oral Reading Fluency Spring; DRA F = Developmental Reading Assessment Fall; DRA S = Developmental Reading Assessment Spring.

Table 2 contains the correlations between both the DORF scores and the DRA scores with the ELA performance level scores. Results indicate statistically significant correlations among the seasonal administrations of each measure as well as between the two screening measures. The fall administration of DORF scores correlate significantly with both the winter and spring administrations of DORF with correlations of r (183) = .849, p < .01 and r (183) = .822, p < .01 respectively. Scores for the winter and spring administrations of DORF correlate significantly as well r (184) = .871, p < .01. Scores for the two administrations of the DRA also correlate significantly with one another r (169) = .660, p <.01.

Results also indicate significant positive correlations between the two different screening measures. Scores from the fall administration of DORF correlate significantly with both the fall and spring administrations of the DRA r (169) = .648, p < .01 and r (183) = .411, p < .01 respectively. Scores from the winter administration of DORF correlate significantly with both the fall and spring administrations of the DRA as well r (169) = .666, p < .01 and r (183) = .438, p < .01 respectively. Finally, scores for the spring administration of DORF correlate significantly with both the fall and spring administrations of the DRA r (169) = .689, p < .01 and r (183) = .502, p < .01 respectively.

Furthermore, results indicate significant positive correlations between both of the curriculum-based measures of reading performance and the ELA performance level scores (p < .01). Scores from the fall administration of the DRA correlate significantly with the ELA performance level scores, r (169) = .404, p < .01. Scores from the spring administration of the DRA also correlate significantly with the ELA performance level scores, r (183) = .355, p < .01. For the DORF measures, scores for the fall and winter administration periods correlate significantly with the ELA performance level scores, r (183) = .302, p < .01 and r (184) = .383, p

< .01 respectively. Scores for the spring administration of DORF also correlate significantly with the ELA performance level scores, r (189) = .402, p < .01.

Table 2

Intercorrelations for Scores on the DORF and DRA and the ELA

Measure	ORF F	ORF W	ORF S	DRA F	DRA S
ELA PL	.302 *	.383 *	.402 *	.404 *	.355 *
ORF F	1	.849 *	.822 *	.648 *	.411 *
ORF W	.849 *	1	.871 *	.666 *	.438 *
ORF S	.822 *	.871 *	1	.689 *	.502 *
DRA F	.648 *	.666 *	.689 *	1	.660 *
DRA S	.411 *	.438 *	.502 *	.660 *	1

Note. ELA PL = English Language Arts Examination Performance Level; ORF F = Oral Reading Fluency Fall; ORF W = Oral Reading Fluency Winter; ORF S = Oral Reading Fluency Spring; DRA F = Developmental Reading Assessment Fall; DRA S = Developmental Reading Assessment Spring.

*$p < .01$.

ROC Curve Analyses

In order to more fully explore the potential predictive nature of DORF and the DRA to ELA performance, a series of Receiver Operating Characteristic (ROC) curves were created that represented the diagnostic accuracy of each screening measure over a range of cut-scores (Streiner & Cairney, 2007). The development of a ROC curve involves plotting the sensitivity against the specificity in order to determine the value of the measure that best estimates performance on the standard measure (in this case the performance level ELA score). The optimum cut-score that represents performance on the ELA as predicted by either DORF or the DRA is the "shoulder" of the curve (i.e., the portion of the curve closest to the upper left corner of the graph). Therefore, an optimal ROC curve would closely follow the vertical axis of the graph to the upper left corner and continue horizontally through the upper portion of the graph. The upper left corner of a ROC curve graph represents a sensitivity of 100 percent and a false-positive rate of 0 percent. However, ROC curves composed of instruments that do not discriminate well would display curves that fall closer to the diagonal between the lower left corner and the upper right corner of the graph (Hintze, Ryan & Stoner, 2003). The diagonal line running from the lower left corner of the graph to the upper right corner is therefore indicative of a screening measure that is completely ineffective at discriminating between two different outcomes (Streiner & Cairney, 2007).

Another statistic described by the ROC curve is the area under the curve (AUC). The AUC is representative of the probability that the screening measure will correctly identify a child as at-risk for failing the ELA who will actually go on to fail the ELA. Therefore, the AUC value gives the probability that the screening measure has accurately identified children as "at- risk." A measure with a larger AUC possesses greater discriminatory ability (i.e., effectiveness) (Streiner

& Cairney, 2007). According to Streiner and Cairney (2007), the following AUC values can be used to determine the accuracy of tests: AUCs between .50 and .70 are considered low, AUCs between .70 and .90 are considered to have moderate accuracy, and measures with AUCs above .90 are considered highly accurate.

For DORF, the predictive validity of the fall, winter, and spring administrations was supported. For the fall administration AUC = .641, $p < .01$ indicating the fall administration of DORF to second grade students is a valid predictor of student performance on the NYS ELA the following school year. For the winter and spring administrations of DORF AUC = .641, $p < .01$ and AUC = .626, $p < .01$ respectively indicating support for the predictive validity of these measures as well. For the DRA, the predictive validity of the measure for both administration periods was also supported. For the fall administration AUC = .664, $p < .01$ and for the spring administration AUC = .619 $p < .01$.

Figures 1, 2, and 3 represent the ROC curves for both assessment tools by time of year such that curves for fall, winter, and spring are represented. Both DORF and the DRA were found to be valid predictors of later student performance on the NYS ELA at each assessment period. Despite the AUC values for both measures falling in the "low" range, the predictive validity of the measures was supported due to the significance level of each measure falling below the .05 cut-off. Thus, the significance level indicted the predictability provided by the screening measure was better than what would be expected by chance.

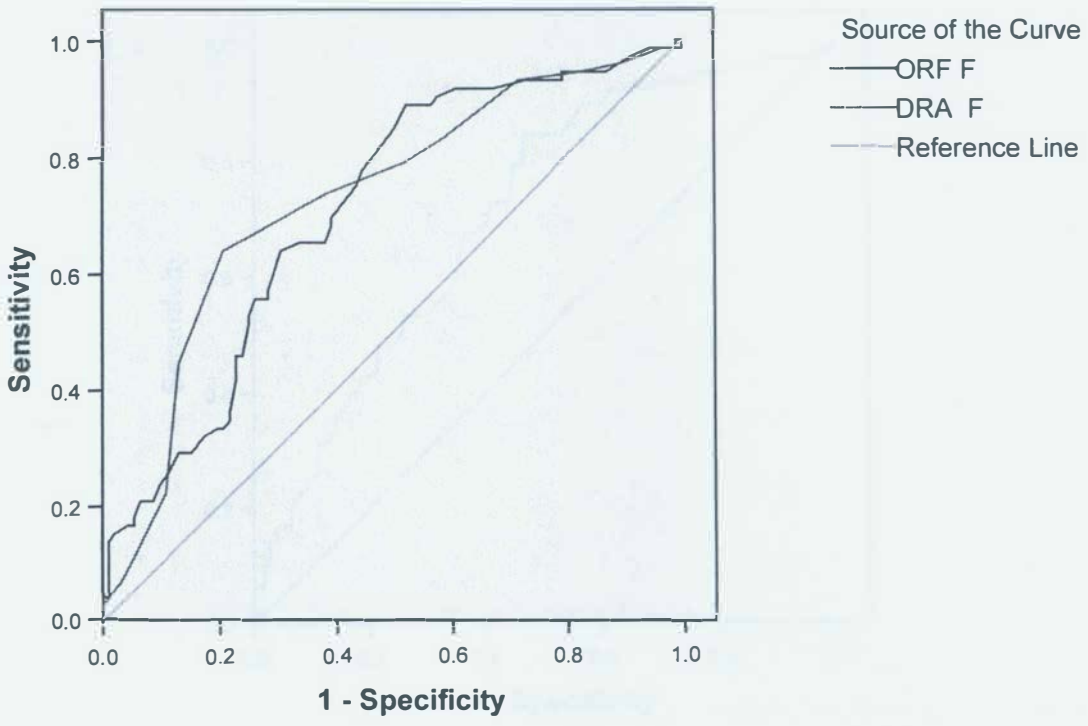

Figure Caption

Figure 1. Receiver Operating Characteristic (ROC) curve of fall screening measures in relation to third grade ELA performance level scores.

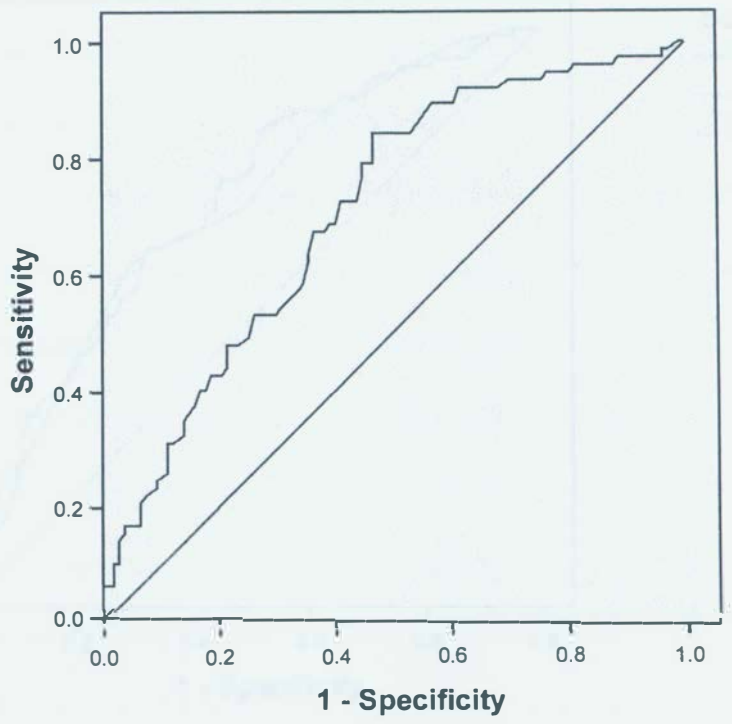

Figure Caption

Figure 2. Receiver Operating Characteristic (ROC) curve of winter screening measure (DORF Winter) in relation to third grade ELA performance level scores.

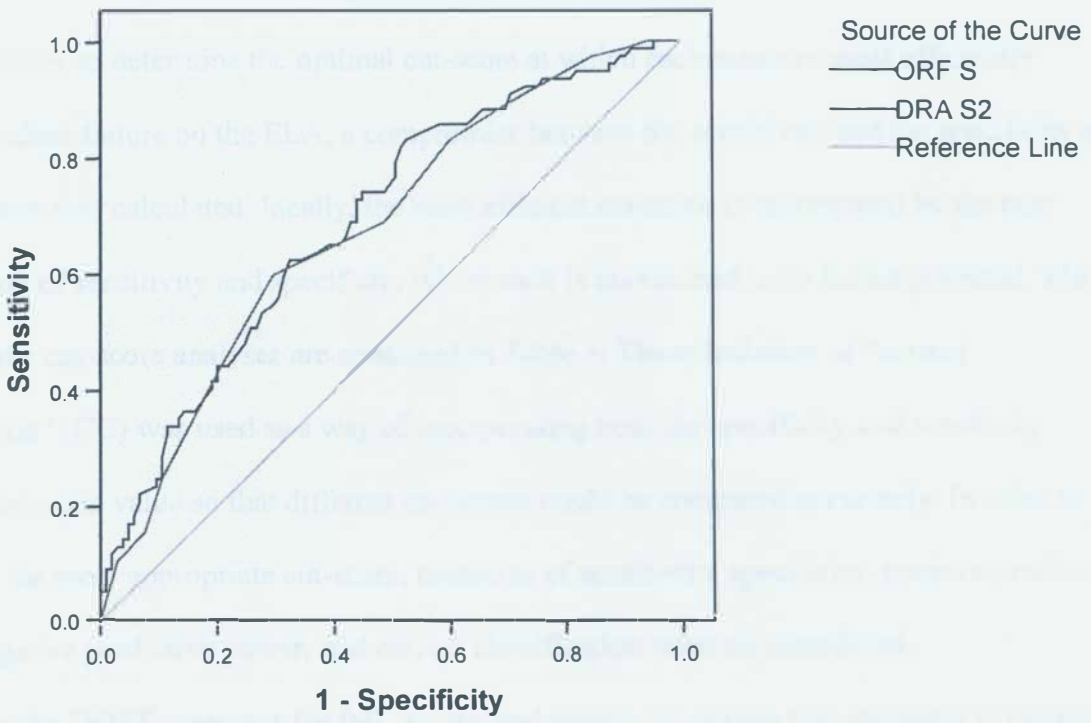

Figure Caption

Figure 3. Receiver Operating Characteristic (ROC) curve of spring screening measures in relation to third grade ELA performance level scores.

Diagnostic Accuracy Analyses

In order to determine the optimal cut-score at which each measure most efficiently predicts student failure on the ELA, a compromise between the sensitivity and the specificity of each measure was calculated. Ideally, the most efficient cut-score is represented by the best combination of sensitivity and specificity where each is maximized to its fullest potential. The results of the cut-score analyses are contained in Table 3. The calculation of "correct classification" (CC) was used as a way of incorporating both the specificity and sensitivity measures into one value so that different cut-scores could be compared accurately. In order to determine the most appropriate cut-score, measures of sensitivity, specificity, positive predictive power, negative predictive power, and correct classification were all considered.

For the DORF measures for fall, winter, and spring, cut-scores that indicated at increased likelihood of failing the ELA were determined to be 45 WRC, 65 WRC, and 90 WRC respectively. For the fall administration, a cut-score of 45 WRC most efficiently identified students likely to fail the ELA with a CC of .65. For the winter administration, a cut-score of 65 WRC was determined most efficient based on the calculated CC of .65. Finally, for the spring administration of DORF, a cut-score of 90 was determined to most efficiently identify students likely to fail the ELA with a CC of .67. For the DRA measures for fall and spring, cut-scores were determined to be levels 12 and 16 respectively. The fall DRA cut-score of 12 resulted in a CC of .72 and the spring DRA cut-score of 16 resulted in a CC of .67.

Table 3

Performance of the DIBELS and DRA over a Range of Cut Score at Each Administration Period

DORF Fall					
ORF F cut score	Sensitivity	Specificity	PPP	NPP	CC
30	.37	.80	.69	.71	.60
35	.47	.72	.67	.65	.65
40	.60	.57	.63	.53	.63
45 *	.70	.56	.66	.53	.65

DORF Winter					
ORF W cut score	Sensitivity	Specificity	PPP	NPP	CC
55	.43	.74	.68	.64	.64
60	.50	.68	.67	.60	.65
65 *	.56	.63	.66	.56	.65
70	.62	.59	.66	.53	.65

DORF Spring					
ORF S cut score	Sensitivity	Specificity	PPP	NPP	CC
80	.55	.61	.64	.54	.67
85	.60	.54	.63	.49	.66
90 *	.70	.49	.64	.45	.67
95	.74	.43	.62	.40	.65

DRA Fall					
DRA F cut score	Sensitivity	Specificity	PPP	NPP	CC
10	.60	.64	.69	.59	.63
12 *	.90	.39	.64	.38	.72
14	.95	.10	.56	.10	.60
16	1.00	.02	.55	.02	.65

DRA Spring					
DRA S cut score	Sensitivity	Specificity	PPP	NPP	CC
12	.25	.82	.66	.65	.51
14	.49	.62	.64	.53	.60
16 *	.80	.38	.64	.36	.67
18	.93	.15	.60	.15	.63

Note. ORF F = Oral Reading Fluency Fall; ORF W = Oral Reading Fluency Winter; ORF S = Oral Reading Fluency Spring; DRA F = Developmental Reading Assessment Fall; DRA S = Developmental Reading Assessment Spring.

* Denotes the cut-score chosen to most efficiently predict passing the ELA.

Comparison of Established District Benchmarks to Cut-scores of the Present Study

The results of comparisons between the district's established benchmark goals and the cut-off scores calculated in the present study are presented in Table 4. Results indicate benchmarks and cut-scores for DORF were similar with scores for the fall being 44 WRC and 45 WRC respectively. The benchmark for the winter administrations of DORF was 68 WRC compared to 65 WRC in the present study. The benchmark for the spring administration of DORF was 90 WRC compared to 90 WRC in the present study. For the DRA, district benchmarks and cut-scores established in the present study differed. The benchmark for the fall administration of the DRA was level 18 compared to a level 12 in the present study. For the spring administration of the DRA, the benchmark was level 28 compared to level 16 in the present study.

Table 4

Comparison of Established Benchmarks for DORF and the DRA with Cut-Scores Established in the Present Study

		Established Benchmarks	Cut-Scores for Present Study
DORF			
	Fall	44 WRC	45 WRC
	Winter	68 WRC	65 WRC
	Spring	90 WRC	90 WRC
DRA			
	Fall	Level 18	Level 12
	Spring	Level 28	Level 16

Note. DORF = Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency; DRA = Developmental Reading Assessment; WRC = Words read correct per minute.

CHAPTER V

Discussion

The purpose of this study was to examine the usefulness of both DORF and the DRA as screening measures to predict later student performance on the ELA. Results indicated significant correlations between DORF and the DRA suggesting these measures are related to one another. They are similar in that they are both meant to assess components of reading development. Although DORF and the DRA vary in how they assess reading skills, there appears to be considerable overlap in what each is measuring.

Relationship of DORF and the DRA to the ELA

Results regarding the predictive utility of these two screening measures on the ELA suggest that both the DORF and the DRA can effectively predict scores on this outcome measure to some extent. DORF and DRA scores for students in second grade exhibited low to moderate correlations with the ELA scores for the same students in third grade. Specific seasonal administration periods of each screening measure did not differ significantly in their predictive validity with regard to the ELA as evidenced by their Pearson correlation coefficients. Furthermore, this study intended to identify which assessment tool, the DORF or the DRA, was more effective as a screener to predict future student performance on the ELA. Given that the differences among the correlations between the two potential screening measures and the ELA were small, the results suggest that both screeners seem to be equally effective at predicting later student performance on the ELA.

The ROC curve data provided an additional source of information on the usefulness of the DORF and DRA as screening measures to predict performance on the ELA. Results indicated both screening measures were valid predictors of later student performance on the ELA based

upon the statistical significance of the AUCs. Taken together, the ROC curve analyses and the descriptive statistics suggest these two screening measures are moderately effective at appropriately predicting student performance on the ELA during the following school year.

Utility of Present District Benchmarks

This study also intended to determine specific cut-scores or risk indicators able to differentiate at-risk students from students not at risk for failing the ELA. Results indicate the established DORF benchmark goals and the cut-scores from the current study are similar suggesting that the current benchmarks can be used as cut-off scores to accurately predict student performance on the ELA within the district included in the study.

The district benchmarks for the DRA are less consistent with the derived cut scores from the current study for both assessment periods. The district benchmarks are much higher than what is actually necessary for a student to likely be considered not at risk for failure of the ELA in third grade. The reason for the discrepancy between the district's benchmarks and the benchmarks determined in the present study may be that the district's benchmarks are based upon assessment at the instructional level on the DRA (i.e., the level at which the student is not reading independently) rather than at the independent level (i.e., the level at which the student is successfully reading on his or her own). In the present study, students were assessed to the independent level. Benchmarks for the independent level would be lower than benchmarks for the instructional level because a student would be more successful reading at the independent level (i.e., students read easier material more successfully). Results suggest that the district may benefit from utilizing the benchmarks established in the present study to identify students when assessing to the independent level with the DRA.

In regard to modifying cut-scores, the selection of appropriate cut-scores is based upon several factors, particularly the type of decision that is to be made. Low stakes decisions can afford a high percentage of false positives; therefore, a relatively liberal cut-score can be used. More conservative cut-scores can be used if the assessor needs to make a more accurate prediction or has fewer assessment resources available. Thus, consideration needs to be given to the types of decisions being made as well as the potential consequences of incorrect decisions (Hintze, Ryan, & Stoner, 2003). In the present study, the district would need to establish cut-scores by balancing the importance of providing additional reading support services to a student who might have ultimately been successful on the ELA without those supports versus the potential detrimental effects of incorrectly identifying a student as not requiring additional support and thus failing to provide that support to a child who actually needs it and would ultimately go on to fail the ELA.

Furthermore, the cut-scores calculated in the present study can be compared with those determined from previous studies correlating R-CBM measures with high-stakes reading achievement tests. Out of the four studies previously discussed in which cut-scores were calculated, two calculated those cut-scores for fourth grade students, one for third grade students, and one for second grade students. For second grade students, Hintze and Silberglitt (2005) determined a cut score of 88 WRC for the spring administration of the ORF measures was able to differentiate between student performance on the MCA. Furthermore, 41 WRC and 71 WRC were determined as cut-scores for the fall and winter administrations of the ORF measures respectively. Results of the current study corroborate the findings of Hintze and Silberglitt (2005) such that similar cut-scores were established. These similar results suggest support for the

creation of performance cut-offs or risk indicators as well as the utility of these particular cut-scores.

Limitations of the Present Study

The current study contained several limitations that could be improved upon by future research. First, due to the archival nature of the database, the researchers did not have information relating to the reading interventions provided by the district to the students during the time of the study. Therefore, the impact these interventions may have had on the study results is not fully known. The predictive validity of a screening measure can be affected by the interventions put in place during the study. As noted by Good et al. (2001), the measurement system has the ability to inform instruction which potentially may lead to changes in instructional programming that can, it turn, bring about changes in student performance if effective teaching strategies are successful (Good at al., 2001). Thus, the use of several screenings throughout the school year enables educators to identify students who are and are not benefiting from interventions that have been put into place in the classroom.

Another limitation regarding assessment fidelity exists because the data was collected by district faculty as opposed to the researchers. Thus, information on assessment fidelity, including interrater reliability values, is not available. Furthermore, the extent of the training of those persons responsible for collecting the data is not fully known.

In regard to assessment with the DRA in particular, students were assessed to the independent level. Data regarding the DRA benchmarks is based upon assessment to the instructional level thus affecting the calculation of the cut-scores. In addition, the district's process for establishing the DRA benchmarks is not fully known. The researchers utilized the benchmarks provided by a district representative as a means of comparison. Furthermore,

information pertaining to the years of experience of each teacher administering the DRA was not known. Since administration and scoring of the DRA relies on teacher judgement, the level of experience of the teacher could influence DRA scores thereby affecting the diagnostic accuracy of the DRA.

Other limitations to generalization of the results of the current study include the lack of information relating to the demographics of the particular sample of students included in the study. District demographic information may accurately represent the demographic characteristics of the present sample. Thus, specific information regarding the use of R-CBM with specific ethnic or gender groups was not obtained.

Implications for Theory

In a prevention-oriented assessment and intervention system, the usefulness of a risk indicator is not solely based on the predictive validity of the measure in relation to a specific outcome measure (Good et al., 2001). The utility of a risk indicator is also based upon the information it can provide prior to any outcome measure. That is, risk indicators serve the equally important role in providing a source of information that can drive instructional changes. Ideally, continued monitoring with measures such as those used in the present study would inform instruction to the degree that original predictions of student performance on outcome measures would no longer be accurate. That is, the overall goal of utilizing risk indicators is that information related to student progress will be provided in a timely manner affording educators the opportunity to make changes in a student's instructional programming that will enable a child who was predicted to be at-risk for failing the outcome measure to be successful and, in turn, continue on the path toward lifelong literacy. Results of the current study provide support for the usefulness of both DORF and the DRA as measures that can provide this information accurately

to educators. Through multiple administrations of both of these measures throughout the early elementary years, information about the risk level of individual students can be obtained prior to administration of the outcome measure in order to improve reading outcomes before failure occurs.

Implications for Practice

For practicing school psychologists, results of the present study have many implications. First, these results support the notion that knowledge of these and other screening measures can provide opportunities for school psychologists to increase their role in consultation and provide assistance in data-based decision making regarding the quality of instruction and the utility of different intervention strategies (McGlinchey & Hixson, 2004). Second, the results emphasize the continued importance of early intervention and primary prevention. These results provide support for a method of both collecting and analyzing data that can be used to identify and assist students in need of increased academic support (Hintze & Silberglitt, 2005). Finally, these results emphasize the importance of setting standards in the school setting, thus providing teachers and other faculty a set of specific scores for goal-setting.

Directions for Future Research

Future research on this topic can aim to extend these results by including high-stakes measures of reading achievement from different states. In addition, future studies may focus on different populations of students to continue to develop research from diverse populations. In order to gain more information regarding the usefulness of this DRA as a screening measure, additional research is needed to evaluate district benchmarks when students are assessed to their instructional level. Extending the research on this topic to other screening measures including

measures used with preschool children will provide more information regarding the relationships between screening and high-stakes assessments.

Although the present study investigated predicting high stakes test scores across years (i.e., second grade screening measures predicting third grade ELA performance), variations of this approach may provide additional information regarding the relationship between screeners and outcome measures. For example, future research could investigate relationships within years (e.g., third grade screening measures predicting third grade test scores) or research might focus on determining which screening period (e.g., fall, winter, or spring) provides the most useful data for predicting performance on an outcome measure. Furthermore, studies that track students' long-term outcomes into the higher grades may be beneficial in identifying additional applications of benchmark or risk indicator development (Good et al., 2001).

Future studies can also focus on the incorporation of cut-scores into districts' policies regarding early intervention and prevention of reading failure (Silberglitt & Hintze, 2005). Finally, future research can focus on alternative sources of information that may assist in the prediction of student success on high-stakes tests. For example, the role of teacher judgment as a predictive tool can be addressed in future studies.

In conclusion, the current study focused on determining whether a relationship exists between the screening measures used and a high-stakes test of reading achievement. Results indicated a significant relationship between the screening measures and the outcome measure, leading to the development of cut-scores for identifying students at risk for not meeting expectations on the state test. These cut-scores or risk indicators were compared to those used by the district and those established from past studies. Results indicated strong relationships between the previously established cut-scores and those established in the present study.

However, district benchmarks for one of the screening measures (the DRA) were deemed inappropriate for accurately identifying at-risk students given the results of this study. Recommendations were made regarding more appropriate cut-scores for the DRA. Given the knowledge of the relationships between screening measures and high-stakes assessments, the goal of reading instruction must focus on the most effective ways of using that knowledge to ensure students receive appropriate support in order to acquire the necessary reading skills needed to function best in society today.

References

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

Beaver, J. (2004). Developmental Reading Assessment Technical Manual. Lebanon, IN: Pearson Learning Group.

Brandt, D. (2001). *Literacy in American lives.* Cambridge, New York: Cambridge University Press.

Burns, M., Griffin, P., Snow, C. (1999). *Starting out right: A Guide to Promoting Children's Reading Success. Specific recommendations from America's leading researchers on how to help children become successful readers.* National Academy of Sciences: National Research Council: Washington D.C.

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

Davis, N., Lindo, E., Compton, D. (2007). Children at risk for reading failure. *Teaching Exceptional Children, 39,* 32-37.

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

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

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

Fuchs, L. & Deno, S. (1981). A comparison of reading placements based on teacher judgment, standardized testing, and curriculum-based assessment. *University of Minnesota.*

Minneapolis, MN.

Fuchs, L., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment. *School Psychology Review, 28,* 659-671.

Fuchs, L., & Fuchs, D. (2004). Determining adequate yearly progress from kindergarten through grade 6 with curriculum-based measurement. *Assessment for Effective Intervention, 29,* 25-37.

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

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

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

Glover, T., & Albers, C. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45,* 117-135.

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

Good, R., Kaminski, R., Simmons, D., & Kameenui, E. (2001). Using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an outcomes-driven model. *Oregon School Study Council Bulletin, 44,* 3-26.

Hasbrouck, J., & Tindal, G. (1992, Spring). Curriculum-based oral reading fluency norms for students in grades 2 through 5. *Teaching Exceptional Children,* 41-44.

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

Hintze, J., 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, 32,* 541-556.

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

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.

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

Lee, J., Grigg, W., & Donahue, P. (2007). *The Nation's Report Card: Reading 2007.* National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. R. Shinn (Ed.), *Curriculum-based*

measurement: Assessing special children (pp. 18-78). New York, NY: Guilford Press.

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

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* U.S. Department of Education. Accessed November 13, 2006, from http://www.ed.gov/pubs/NatAtRisk.

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

New York State Department of Education. (2005). *English language arts core curriculum.* Retrieved November 11, 2006 from http://www.emsc.nysed.gov/ ciai/ela/ elacore.

New York State Department of Education. (2006). *New York State testing program 2006: English language arts, grades 3-8* (Technical Report). Monterey, CA: CTB/McGraw-Hill.

No Child Left Behind Act of 2001. Pub. L. No. 107-110.

Pearson Learning Group. (2003). Developmental Reading Assessment (DRA) K-8 Technical Manual.

Shinn, M., Good, R. & Knutson, N., Tilly, W. & Collins, V. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21,* 459-480.

Sibley, D., Biwer, B., & Hesch, A. (2001). Establishing curriculum-based oral reading fluency performance standards to predict success on local and state tests in reading. [Electronic version]. Unpublished data, Presented at the Annual Meeting of the National Association of School Psychologists, Washington D.C.

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

Sloat, E., Beswick, J., & Willms, D. (2007). Using early literacy monitoring to prevent reading failure. *Phi Delta Kappan, 88,* 523-529.

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

Streiner, D. & Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristics curves. *The Canadian Journal of Psychiatry, 52,* 121-128.

U.S. Department of Education. (2002). *The facts about...Reading First.* Retrieved November 13, 2006, from http://www.nochildleftbehind.gov

U.S. Department of Education. (2004). *A guide to education and No Child Left Behind.* Retrieved October 13, 2006 from http://www.ed.gov/print/nclb/overview/intro/guide/guide.html.