2008

Establishing universal screening risk indicators using reading curriculum-based measurement and the developmental reading assessment

Adrienne M. Zonneville

Running head: ESTABLISHING UNIVERSAL SCREENING RISK INDICATORS

Establishing Universal Screening Risk Indicators using Reading Curriculum-Based Measurement and the Developmental Reading Assessment

Adrienne M. Zonneville

Rochester Institute of Technology

School Psychology Department

First Reader: Dr. Suzanne Graney

Second Reader: Dr. Vincent Pandolfi

Abstract

The current study examined the abilities of Dynamic Indicator of Basic Early Literacy Skills Oral Reading Fluency (DORF) and the Developmental Reading Assessment (DRA) to be used as screening measures for predicting performance on the yearly state reading assessment in New York. Data from 194 students' performance on DORF, DRA, and the English Language Arts (ELA) assessment were examined from an existing data base. Participants' scores on the DORF and the DRA in third grade were compared to their performance on the fourth grade ELA. Screening cut scores were developed for each benchmarking period to assist in identifying students who need additional reading support and intervention. Patterns of correlations between the screening measures and the ELA were examined. Risk indicators were determined by examining receiver operating characteristic (ROC) curves and by creating screening outcome matrices. Results support DORF risk indicators 70 words read correct (WRC) or fewer in the fall, 80 WRC or fewer in the winter, and 100 WRC or fewer in the spring and DRA risk indicators of 14 or fewer in the fall and 16 or fewer in the spring. Results further support the use of the DORF and DRA as screening measures for identifying students at risk for low reading skills and failing state assessments.

Chapter I

Statement of the Problem

Illiteracy is a nationwide problem. The National Assessment of Educational Progress (NAEP) indicates that approximately 40 percent of students cannot read at a basic level (U.S. Department of Education, 2002). When considering subgroups of students, almost 70 percent of low-income fourth graders and almost 50 percent of students living in urban areas cannot read at a basic level (U.S. Department of Education). It is estimated that 10 million children, or 1 child in 5, experience significant difficulties learning to read at a level in which they can use reading to learn (Lyon, 1998).

Reading is a critical skill that all children need to develop in order to succeed in today's schools. Children who are unable to read are largely at-risk for school failure and future occupational and vocational failure. Successful reading is vital to success in all realms of our society (Lyon, 1998). Illiteracy is related to high-school dropout rates, incarceration, lack of civic awareness, poor health maintenance, and poverty. It is therefore essential that schools reduce the prevalence of reading failure (Fuchs & Fuchs, 1999).

Federal Initiatives

Current federal initiatives such as the No Child Left Behind Act of 2001 (Public Law 107-110; NCLB) have increased demands for early identification and intervention and increased state and district accountability. NCLB was designed to improve student achievement. It represents a federal effort to support early elementary and secondary education in the United States. The Act places an increased emphasis on reading and implementing practices that have been clearly demonstrated to be effective through rigorous scientific research. NCLB is aimed at helping all students meet high academic standards by requiring all states to create annual

assessments in grades three through eight in reading and math (Allington, 2005). Individual students who do not pass the assessment must receive remedial instruction. Currently, each school district must decide whether students who fail the exam can be promoted to the next grade level (New York State Education Department).

Reading First is a competitive grant program authorized under NCLB as a nationwide effort to improve the reading skills of students in kindergarten through third grade. The program provides grants to school districts that submit an approved application (U.S. Department of Education, 2002). Reading First is designed to help states, districts, and communities identify and adopt scientifically based reading programs and ensure all classroom teachers for grades kindergarten through third can identify children at risk for reading failure and provide effective early instruction and intervention (Kauerz, 2002).

To ensure students receive appropriate reading instruction, Reading First requires the use of validated and reliable screening, diagnostic, progress monitoring, and classroom-based reading assessments (Sopko, 2002). Screening and diagnostic tools, such as Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency (DORF) and the Developmental Reading Assessment (DRA), are used to measure student reading and monitor progress (Sopko; U.S. Department of Education, 2002). Reading programs must include the empirically validated five essential components of reading instruction according to the National Reading Panel: phonemic awareness, phonics, vocabulary development, reading fluency, and reading comprehension (National Reading Panel, 2000). These big ideas in reading provide a foundation for reading success (National Reading Panel, 2000).

Current educational reform efforts have emphasized increases in student achievement as demonstrated by scores on statewide achievement tests. Performance on state assessments

influences much of the educational decision making for school districts (McGlinchey & Hixson, 2004). High-stakes assessments have focused national attention, effort, and resources on reading outcomes (Good, Kaminski, Simmons, & Kame'enui, 2001). They are generally driven by increased efforts at accountability and a need to measure student progress relative to the instructional curriculum. Standardized high-stakes tests provide annual information about students' global standing. Reading skills are sampled across several years of curriculum, providing teachers with little diagnostic information (Fuchs & Fuchs, 1999).

State assessments are inappropriate for early identification of students at risk for reading difficulties (Fuchs & Fuchs, 1999). They often are administered on an infrequent basis, providing teachers with limited information about students' progress toward mastering academic benchmarks. They fail to provide teachers with diagnostic information related to student attainment of specific instructional goals. Yearly state assessments provide summative information identifying children only after they have not met the standards, at a time when it is often too late to modify reading instruction (Good, Simmons, & Kame'enui, 2001). Standardized tests are not able to provide accurate, comprehensive, or detailed results (Neill, 2006) about students' academic proficiency (Good et al.). They are technically inadequate for making decisions about individual students and are not useful for making instructional decisions (Deno, 1985).

Adequate assessments inform instruction, not simply describe how students are performing. Assessment data need to inform referral, screening, classification and entitlement, instructional planning, and progress monitoring decisions. Students' academic progress needs to be closely monitored and screened through other measurement systems (Sibley, Biwer, & Hesch, 2001; Deno, 2003). A poor match often exists between the local curriculum and the test content.

This creates problems when attempting to interpret the results of state assessments. Many districts are now changing curriculum content and revising instructional approaches as a result of pressures to report scores that reflect increases in student performance. In essence, teachers are beginning to "teach to the test" (Neill, 2006).

Assessment in a Prevention Oriented Framework

A prevention-oriented, school based assessment and intervention system can be used to predict reading success and difficulty early and to inform instruction responsively. If a universal screening assessment system is in place early, it can be used to signal reading difficulty and prevent early reading risk from becoming reading failure (Good et al., 2001). Such an assessment system must be able to measure and monitor changes in student performance over time. Assessment systems for educational accountability and prevention must document and account for growth in foundational reading skills, predict success or failure on criterion measures of performance, and provide an instructional goal that if met will prevent reading failure and promote reading success (Good et al.). Using an assessment system focused on prevention will match students' needs to the instructional support necessary to help them achieve in reading before a pattern of reading difficulty and failure is established (Good, et al.).

Universal screening is critical to providing early school-based prevention and intervention services for students at risk for reading difficulties. Early screening is a vital aspect in the provision of targeted prevention and intervention services (Glover & Albers, 2007). School-based universal screening is typically conducted with all students in a classroom to identify those at risk for academic or behavioral difficulties. Once these students are identified, they could receive specific instruction and intervention (Glover & Albers). In recent years, it has

become more common to screen all students and identify those who are at risk for academic failure. Students' individual performance can be compared to that of a group (Deno, 2003).

Progress monitoring is vital to a prevention oriented system. Progress monitoring is the scientifically based practice (Kim-Sung, 2006) of assessing students' academic performance on a regular basis to determine whether students are making progress in their current instructional program and to build more effective intervention programs for children who are inadequately benefiting from typical instruction (Fuchs & Fuchs, 1993). Benefits of progress monitoring include accelerated learning, more informed instructional decisions, and higher expectations for students by teachers (Kim-Sung).

Reading Curriculum-Based Measurement

Measurement of student achievement is essential to evaluating the success of educational programs in a prevention-oriented model. Reading curriculum-based measurement (R-CBM) represents an effort to decrease the separation between measurement and instruction. It was originally developed for use by special education teachers to evaluate student progress and instructional effectiveness (Deno, 1985). Well-designed classroom-based assessments can provide a richer, more consistent indicator of a child's performance compared to year-end assessments. Curriculum assessments provide information that can be used to increase the validity, diagnostic capacity, and the ability to assess progress toward attainment of meaningful standards (Deno).

R-CBM is standardized. It provides teachers with a reliable, valid, and efficient procedure for obtaining ongoing performance with which to evaluate instructional programs. It is easily understood and inexpensive (Deno, 1985). It has been established through research to be a non-biased assessment (Hintze, Callahan, Matthews, Williams, & Tobin, 2002). Testing

methods and testing content remain constant so that progress can be monitored systematically over time. R-CBM samples many dimensions of curriculum throughout the year (Fuchs & Fuchs, 1999). Reading difficulty can be distinguished by comparing performance levels between individuals. Research-based benchmarks have been created for the fall, winter, and spring of each grade level to specify the minimum performance levels associated with reading success (Fuchs, Fuchs, & Compton, 2004). Many alternate test forms are provided, permitting repeated measures over time (Fuchs, Fuchs, Hosp, & Jenkins, 2001). Scores represent accurate and meaningful estimates of reading competence. A CBM screening assessment allows schools to distinguish students that are not on track to be proficient readers from an early age and enables educators to monitor the effectiveness of interventions designed to address the specific needs of these students (Silberglitt & Hintze, 2005).

R-CBM uses oral reading fluency (ORF) as an indicator for measuring reading achievement. Oral reading fluency is the translation of text with speed and accuracy. It directly measures phonological segmentation, decoding skill, and rapid word recognition (Fuchs et al., 2001). Oral reading fluency develops gradually over the elementary school years and can be indexed as words read correctly per minute (Fuchs et al.). The number of words read aloud correctly and incorrectly in one minute from a basal text has been shown to discriminate growth in reading proficiency throughout the elementary school years (Deno, 1985). Words read correctly per minute is valid, reliable, simple and quick, inexpensive, and easily understood, can be given often, and is sensitive to growth over short periods of time. Research has found strong support for the efficiency of oral reading fluency among general education and special education populations as a measure of reading proficiency and comprehension (Jenkins, Fuchs, Broek, Espin, & Deno, 2003; Shinn, Good, Knutson, Tilly, & Collins, 1992).

R-CBM measures can be used with all children in a school to assist in the detection of reading problems early. These methods can be easily integrated into a problem-solving model. (Ardoin, Witt, Suldo, Connell, Koenig, Resetar, et al., 2004) R-CBM can be used to identify students in need of intensive instruction and intervention. Traditionally, this screening process has been achieved through the administration of standardized assessments and the use of teacher judgment (Madelaine & Wheldall, 2004). Previous research has provided evidence that teachers are not always adequate referral sources. Accurate screening sources must be stable across children, settings, and teachers (VanDerHeyden & Witt, 2005). R-CBM offers an appropriate alternative to teacher judgment and standardized assessments for the practice of initial identification of students at risk of reading problems (Madelaine & Wheldall).

Relationship between R-CBM and High-Stakes Testing

Past research conducted in several states, including Oregon, Illinois, Washington, and Michigan has examined the use of R-CBM for predicting performance on state reading assessments (Crawford, Tindal, & Stieber, 2001; Good et al.; 2001; Hintze & Silberglitt, 2005; McGlinchey & Hixson, 2004; Sibley et al.; 2001; Silberglitt & Hintze, 2005; Stage & Jacobson, 2001). Studies have focused primarily on establishing benchmark scores that would predict passing state assessments. These studies have concluded that a moderate to strong relationship exists between student performance on oral reading fluency tasks and student performance on the state assessments evaluated. Benchmark goals have been created and applied in various geographical locations that were predictive of continued progress and desired outcomes such as success on the state assessments. Past research supports that a significant correlation exists between R-CBM and specific state assessments.

The Developmental Reading Assessment

The Developmental Reading Assessment (DRA) is a literature-based instructional reading and assessment program used to help teachers assess and document student reading performance over time. The DRA was developed by the Upper Arlington School District to identify students at risk for reading failure. Districts and states can use the DRA to monitor and evaluate the reading skills and development of students in kindergarten through eighth grade. The DRA is used in more than thirty thousand classrooms across the United States (Pearson Learning Group, 2003).

The DRA manual (Pearson Learning Group, 2003) states that the DRA is designed to inform and shape instruction. It can be used to assess the level at which students can read text independently, and to assess students strengths and weaknesses in relation to engagement, oral reading fluency, and comprehension. The DRA is designed to measure students' reading ability and growth to help teachers identify individual student needs (Beaver & Carter, 2003).

The DRA can be administered annually or semiannually. It can be administered more frequently with struggling readers to monitor progress. It is administered in a one-on-one, or conference, format between student and teacher. Results from the DRA can be used to inform instructional interventions and strategies for students at risk of reading failure (Pearson Learning Group, 2003). Current scientific research is limited with respect to the psychometrics of the DRA and its ability to predict student performance on state assessments.

Purpose of Study

Past research has demonstrated the sensitivity of R-CBM as a measure of progress over time as well as a dynamic indicator of overall reading growth and development. The purpose of this study was to replicate and extend previous research that has examined the relationship

between R-CBM and high stakes testing and identify risk indicators that can be used for universal screening assessments. Furthermore, no previous research has evaluated the relationship between the DRA and high stakes assessments. This study examined student performance on R-CBM, the DRA, and the mandated state reading assessment in New York. Screening cut-scores were identified and risk indicators were developed. These risk indicators can be used to identify students at risk for low reading skills and in need of additional reading supports and interventions. They can also be used to set goals for students receiving interventions. Students not likely to achieve satisfactory scores on the state reading assessment can be identified and reading supports can be given earlier in the school year instead of waiting until the test is administered and scored.

The current study includes students from an urban school district in Western New York. These students represent a larger sample of students and a more diverse student population than many of the previous studies. This study differed from the previous research in that it examined how student performance in third grade on R-CBM and the DRA related to their fourth grade performance on the ELA exam and established risk indicators to predict which students were likely at risk for low reading skills and likely to not experience success on the ELA. No known previous research has examined the relationship between the DRA and high-stakes assessments or empirically evaluated DRA benchmarks for predicting reading performance.

The risk indicators developed in the current study can help teachers identify students at risk for reading difficulties earlier and implement interventions sooner to assist these students and help ensure that they experience reading success. Cut-scores for third grade reading performance on R-CBM and the DRA can assist in identifying students in need of extra support or curriculum adjustments. Implementing a screening assessment for predicting future

performance on the ELA fits into the problem solving model and a prevention oriented assessment framework which focuses on early intervention and prevention of reading failure.

The current study will address the following research questions:

1. To what degree does student performance on R-CBM and DRA screening measures in third grade correlate with fourth grade ELA performance?
2. What score on R-CBM in third grade can be used to establish a screening risk indicator that can be used to identify students not likely to achieve satisfactory scores on the fourth grade ELA exam?
3. What score on the DRA in third grade can be used to establish a screening risk indicator that can be used to identify students not likely to achieve satisfactory scores on the fourth grade ELA exam?

Chapter II

Literature Review

The nationwide problem of illiteracy and the Federal initiatives that have been implemented to ameliorate the problem were reviewed in the previous chapter. The imperative for implementing formative and screening assessments, such as R-CBM and the DRA, within schools was reviewed. Predicting which students are at-risk for experiencing reading difficulties and failure on year-end summative assessments allows schools to intervene and implement interventions sooner. The use of R-CBM and the DRA as measures that can be used to identify students at-risk for reading failure and monitor student progress toward meeting goals was highlighted. Using formative assessment measures, like R-CBM and the DRA, as screening devices was proposed as a method for predicting which students are at risk for low reading skills and for predicting student performance on year-end state reading assessments. Primary research articles supporting the use of R-CBM and the DRA and the use of formative assessment for screening reading problems and identifying students at risk for performing poorly on high stakes assessments will be reviewed in the present chapter.

Reading Curriculum-Based Measurement

The previous chapter provided a general overview of R-CBM. R-CBM provides a method by which student reading performance can be screened and progress can be continuously monitored which aids teachers in making formative decisions (Wiley & Deno, 2005). Formative teaching systems can be useful for indexing individual student growth over time. Decision making regarding student progress can be considered throughout the year and not just end-of-the-year summative decisions (Fuchs & Fuchs, 1993).

Stanley Deno and colleagues at the University of Minnesota were interested in the idea of formative evaluation and sought to decrease the separation between measurement and instruction. Deno and colleagues developed measurement and evaluation procedures for reading, written expression, and spelling that could originally be used by special education teachers to routinely measure student achievement and to make decisions regarding when and how to modify students' instructional programs. This measurement and evaluation system was developed and tested from 1977-1983 (Deno, 1985) with the goal being to test the effectiveness of a special education intervention model called data-based program modification, based on the idea that special education teachers could use repeated measurement data to evaluate instruction (Deno, 2003). Results demonstrated that teachers were more effective when they employed such a system. The measurement procedures resulting form this research are now referred to as curriculum based measurement (Deno, 1985).

When developing a system of curriculum based measurement, Deno and colleagues (1985) believed that curriculum based measures of reading needed to relate to growth in reading comprehension. The research group wanted to establish the reliability and validity of various reading measurement procedures including cloze (supplying words deleted from text), word meaning, and reading aloud tasks. Results of criterion validity studies supported the findings that curriculum based measures were highly correlated with performance on standardized, norm-referenced tests except for the word meaning task. A close relationship was found between reading aloud from text and comprehension scores. Reading aloud from text was found to reliably differentiate between students participating in special education programs and those not in such programs, strengthening the criterion related validity of the measure (Deno, Martson, Shinn, & Tindal, 1983, as cited in Deno, 1985). When cross-sectional analysis of developmental

trends were researched from elementary students across the United States, results indicated that the number of words read correctly and incorrectly from a basal text reliably and validly distinguished growth in reading proficiency (Deno, 1985).

In 1989, Martson reviewed the existing research on CBM. At that time, it was viewed primarily as a progress monitoring tool for elementary students in special education. Initial studies had concluded that listening to students read aloud for one-minute from a basal reader was a valid measure of reading skill; correlation coefficients ranged from .73 to .91 with criterion tests of reading, with most above .80. Later studies found correlation coefficients between oral reading and different published measures of global reading skills ranged from .63 to .90, with most above .80. These studies supported the criterion-related validity of curriculum-based reading measures as a predictor of overall reading skill. Reading aloud from passages, median correlation of .84, has been found to be more highly related to performance on basal mastery tests than was reading from word lists, median correlation of .76. A median correlation of .86 has been found between reading fluency measures and teacher judgments of student reading skills (Martson, 1989).

Test-retest reliability estimates were examined using intervals of 1 to 10 weeks. Reliability coefficients ranged from .82 to .97, with most coefficients above .90. Alternate-form estimates ranged from .84 to .96, with most coefficients above .90. An inter-rater agreement coefficient was found to be .99 (Martson, 1989).

Many additional studies have examined the validity of R-CBM. Overall, the findings have provided strong support for oral reading fluency as a reliable and valid measure of students' general reading skills. Most correlations in these studies have been found to be around .80. R-

CBM has also demonstrated discriminant validity, longitudinal change, sensitivity to changes in reading programs, and treatment validity (Deno, 1985; Shinn et. al., 1992; Wayman et al., 2007).

To address concerns that oral reading fluency measures only decoding skills rather than higher order reading skills such as comprehension, Shinn et al. (1992) investigated the contribution of R-CBM to the reading process from a theoretical perspective. Confirmatory factor analysis was used to examine the relationship of reading aloud to decoding, fluency, and comprehension skills for students in grades three and five. A total of 238 third and fifth grade students from a predominately white public school were administered eight reading measures: two were measures traditionally used to assess decoding skills, four were traditionally used to assess comprehension skills, and two measures assessed oral reading fluency (Shinn et al.).

In third grade, the study found that a one-factor model, labeled "reading competence," was validated with all reading skills making sufficient contributions. R-CBM correlated highly (.88 and .90) with the reading competence construct. A two-factor model including decoding and comprehension as separate but related factors was validated for fifth graders, with reading aloud loading on the decoding factor. Reading decoding was strongly related to reading comprehension (.83). The relationship between oral reading fluency and comprehension was also high. The nature of the relationship between reading aloud and reading proficiency changed with age. The effectiveness of oral reading fluency as a measure of reading proficiency and comprehension was strongly supported regardless of the factor model employed. This study lends support for R-CBM as a measure of general reading achievement, including comprehension, for third grade students (Shinn et al., 1992).

In recent years, R-CBM research has examined racial/ethnic and gender bias on curriculum-based measurement and has produced mixed results. In a study conducted by

Kranzler, Miller, and Jordan (1999), R-CBM was not found to be an unbiased indicator of current reading comprehension for students in grades 4 and 5. Although no evidence of bias was found in grades 2 and 3, a bias for Caucasian and African American students was evidenced in grades 4 and 5. R-CBM reading performance overestimated the reading comprehension abilities of African American students and underestimated the reading comprehension abilities of Caucasian students in the upper elementary grades. In grade 5, R-CBM was found to overestimate the reading comprehension of girls and underestimate the reading comprehension of boys (Kranzler, Miller, & Jordan, 1999).

Hintze et al. (2002) sought to replicate and extended the work of Kranzler et al. (1999). The researchers evaluated the differential predictive bias of R-CBM across African American and Caucasian students in second through fifth grade. Results suggest that CBM was not biased with respect to SES or ethnicity. CBM oral reading scores predicted the same or similar criterion reading comprehension scores regardless of the child's ethnicity. No bias was evidenced across groups with respect to the student's age, R-CBM abilities, SES, or ethnicity. Results of this study suggest that R-CBM is a sensitive form of assessment in curriculum for African American and Caucasian elementary students. Although future studies are needed, based on this study, R-CBM appears to be a sensitive measure of reading achievement for both African American and Caucasian elementary-age students (Hintze et al.).

Teachers can use R-CBM scores as universal screening measures, as a tool for monitoring students' development of academic progress, and to improve instructional programs. Students experiencing reading difficulties in general education relative to same age peers and instructional standards can be identified. Teachers can use R-CBM data to distinguish which skills in the curriculum require additional instruction and which students are experiencing

problems (Fuchs & Fuchs, 2001). Research has demonstrated the usefulness of R-CBM in identifying children for special education, establishing and monitoring progress toward IEP goals, monitoring progress in remedial programs, and designing instruction (Deno, 1985; Fuchs & Fuchs, 2001). R-CBM is also valuable for evaluating the effectiveness of interventions (VanDerHeyden & Witt, 2005).

R-CBM presents a number of advantages over traditional norm-referenced tests. R-CBM focuses on repeated measurement to monitor student progress, is sensitive to change in performance over time, and provides reliable feedback on the effectiveness of instructional interventions. R-CBM provides relevance to instruction by assessing performance using the same materials used in the classroom or by using pre-designed grade level probes (Sibley et al., 2001).

VanDerHeyden and Witt (2005) examined the predictive accuracy of a problem-solving model of assessment which included universal screening and teacher referral. The researches examined the degree to which universal screening and teacher referral correctly identified male and female children, children of minority and non-minority ethnicity, and children in high-achieving and low-achieving classrooms. The research was found to strongly support the use of R-CBM to accurately identify students who were in need of academic support within the general education setting. Curriculum-based measurement was found to be a stable identification source for screening across varying contexts, such as race and gender. In this study, universal screening surpassed an accuracy rate of 93%, whereas teacher referral did not. Universal screening using curriculum-based assessment in the general education setting was used to accurately identify children who were in need of academic assistance (VanDerHeyden & Witt).

Establishing Risk Indicators

School personnel can use R-CBM to inform them when students may not have a high probability of attaining desired outcomes on state and local assessments. Procedures for R-CBM can assist in the identification of students demonstrating reading difficulty and in need of additional reading supports. In order for this to occur, it is necessary to establish risk indicators for R-CBM that are linked to high stakes tests (Sibley et al., 2001). Using risk indicators linked to statewide assessments provides a consistent set of rigorous criteria for judging student performance. Students performing below a given cut-score are likely at risk for reading difficulties and highly probable of experiencing failure on the state test. By using a systematic method for establishing cut-scores at all grades and benchmark periods, educators can easily apply the concepts of formative assessment to evaluate the progress of an individual or group of students. A consistent set of cut-scores will allow for regular, frequent, and valid measurement to a common outcome. Cut-scores can be used during a screening assessment to guide which students are in need of additional reading supports. Interventions can be implemented and student progress can be monitored frequently using R-CBM (Silberglitt & Hintze, 2005; Good et al., 2001).

Relationship between Curriculum-Based Measures and High Stakes Testing

Recent research has expanded on the use of R-CBM beyond classification determination and instructional decision making and has examined how it can be used for making accountability decisions and predicting performance on state assessments. Establishing which students are at risk for low reading skills and failing state tests has been highlighted in the literature. Early studies focused on establishing benchmark scores that would predict passage or failure on state reading assessments. Crawford, Tindal, and Stieber (2001) examined which

levels of oral reading rate in second and third grade best predicted students' successful performance on the Oregon State Assessment (OSA) in third grade. A direct relationship was found between R-CBM oral reading and the likelihood of passing the OSA. Of the students reading at the third and fourth quartiles, 81% passed the statewide assessment. For third grade students, 119 words read correct per minute (WRC) was the critical rate needed to pass the statewide reading test. Of the students reading at least 72 WRC in second grade, 100% passed the statewide reading test in third grade. The use of CBM as a measurement tool for providing information about students' current and future performance was supported (Crawford et al.).

Good et al. (2001) explored a range of indicators of initial early literacy skills to predict emerging reading outcomes as well as performance on the OSA. Like the previous study, the relationship between R-CBM oral reading fluency and student performance on high-stakes reading assessments was explored. Benchmark goals were created by examining the level of proficiency on R-CBM oral reading fluency that predicts successful performance and failure on the OSA (Good et al.)

The usefulness of benchmark goals was strongly supported. Students achieving early benchmark goals were likely to attain subsequent literacy benchmark goals. The first-grade outcomes were predictive of continued progress in second grade and consistent with desired second-grade outcomes. The first grade benchmark goal of 40 WRC on CBM oral reading fluency had utility as a goal that predicted continued reading progress; 97% of students attaining the first-grade benchmark went on to achieve the second-grade benchmark. Of the 51 students reading below 10 WRC in the spring of first grade, none went on to achieve the second-grade benchmark, indicating that a score below 10 appeared to have utility as a level in which intensive instructional support was needed. A second grade benchmark of 90 WRC was identified. Based

on this study, the single best way to increase second-grade reading outcomes is to attain the spring of first grade benchmark goal (Good et al., 2001).

Of students attaining the spring of third grade benchmark, 96% went on to experience success on the OSA. Students reading third grade material at 110 WRC or better were likely to meet expectations on the OSA. Students scoring below 70 WRC were unlikely to meet expectations on the OSA; indicating that these students need intensive instructional support. Results of this study support oral reading fluency as an important foundation for reading competence. Its use as a screening measure for identifying students in need of additional reading supports and interventions was also supported as well as its use in predicting future reading performance (Good et al., 2001).

Sibley et al. (2001) applied the spring oral reading fluency benchmarks previously established in Oregon to student performance data for grades three through five from a suburban school district in Illinois. A significant correlation was demonstrated between R-CBM oral reading fluency benchmarks for second, third, and fourth grade students and performance on the Illinois Standards Achievement Test (ISAT). This study proposed fourth grade oral reading fluency benchmarks linked to fifth grade performance on high stakes assessments. The use of R-CBM performance data to identify students that may not meet standards on state assessments and implement appropriate interventions was further supported. The usefulness of the spring global fluency based performance standards found by Good et al. (2001) for predicting student performance on subsequent high stakes achievement measures was supported.

Later studies exploring the relationship between R-CBM and high stakes assessments have reported diagnostic efficiency statistics, including sensitivity, specificity, positive predictive power, and negative predictive power. Sensitivity refers to the percentage of students below a

cut score that went on to fail the state assessment. Specificity refers to the percentage of students above a cut score that later passed the state assessment. Positive predictive power reflects the probability that a student with a score below the cut score will truly fail the test. Negative predictive power reflects the probability that a student performing above a cut score will truly pass a test (Waymann et. al., 2007). Diagnostic efficiency statistics have been fairly consistent among studies. Sensitivity values have ranged from 65 percent to 76 percent and specificity values have ranged from 74 percent to 82 percent. The use of R-CBM has been established to significantly add to positive and negative predictive power above base rates of prediction (Stage & Jacobson, 2001; McGlinchey & Hixson, 2004; Silberglitt & Hintze, 2005; Hintze & Silberglitt, 2005).

Stage and Jacobson (2001) examined how the previous research on R-CBM applied to the fourth grade Washington Assessment of Student Learning (WASL). Student performance on fourth grade fall, winter, and spring benchmark assessments as well as student growth between each benchmarking period was assessed. Results indicated that the level of each student's performance on oral reading fluency benchmarking better predicted WASL reading performance than the amount of growth in oral reading fluency across the school year. Goals were developed to predict passage and failure on the WASL based on benchmark data. Growth curve analyses and three analyses of variance were conducted to determine the cut score that best predicted passage on the WASL. Diagnostic efficiency statistics were used to accurately identify the students most likely to fail the WASL reading test using their ORF scores. The probability of correctly predicting who would pass the WASL based on the September cut-score was .90 (negative predictive power). The probability of predicting who would fail the WASL based on the September ORF cut-score was .41 (positive predictive power). The observed percent of

students correctly classified was 73.9 percent. September oral reading fluency cut-scores increased the predictive power of failure and success on the WASL by 30 percent over base rate levels. Results support the use of R-CBM for improving the prediction of state test performance, permitting early identification and intervention (Stage & Jacobson).

McGlinchey and Hixson (2004) replicated the previous study by Stage and Jacobson (2001) across 8 years, with a larger sample, across a more diverse student population, and with the Michigan Educational Assessment Program (MEAP). Fourth grade students were administered reading passages two weeks prior to the administration of the MEAP. Similar to the results of Stage and Jacobsen, oral reading fluency was found to improve the prediction of performance on the state fourth-grade reading assessment above that based on the base rates of passing and failing. Results indicate a moderately strong relationship between oral reading rates and MEAP performance. Using 100 WRC as the cut-score, the percentage of students reading at or above the cut-score who later passed the MEAP was 74 percent. The sensitivity of the cut-score for identifying students who did not achieve satisfactory scores on the MEAP was 54 percent. The probability of correctly classifying students who scored below satisfactory was 77 percent while the probability of correctly identifying students who passed the MEAP was 72 percent. Overall, a cut score of 100 WRC resulted in correct classification of 74 percent of students (McGlinchey & Hixson, 2004).

Silberglitt and Hintze (2005) used alternative statistical methods to create benchmarks that could accurately predict student success on the Minnesota Comprehensive Assessment (MCA). Four statistical methods for establishing cut-scores were investigated. Methods were compared on the basis of three factors: "(a) the diagnostic accuracy of the cut scores generated using each method; (b) a visual analysis of how the cut scores compared to those established by

previous research; and (c) an analysis of the appropriateness of each method for the data set given (Silberglitt & Hintze, 2005, p. 311)." The first method used was discriminant analysis. Students were classified into two groups, those that did reach grade level proficiency on the MCA and those that did not. An equipercentile method was used to equate the percentage of students below an identified score on the MCA with that percentage on R-CBM. Logistic regression was used to calculate the probability that a person passed or failed the MCA based on his or her R-CBM score. The fourth method, receiver operating characteristic (ROC) curve analysis, was used to plot the specificity and sensitivity of R-CBM for all possible values of the cut score and was used to determine which cut scores yielded the strongest diagnostic accuracy (Silberglitt & Hintze).

Receiver operating characteristic (ROC) curve analysis provided the most flexibility in establishing desired levels of diagnostic accuracy. Discriminant analysis, logistic regression, and equipercentiles were used to determine a range of possible cut scores. The four methods resulted in cut scores differing as much as 11 WRC, generating a slight effect on overall diagnostic accuracy. ROC curves consistently yielded the highest levels of negative predictive power and sensitivity. Logistic regression consistently identified cut scores with the highest level of diagnostic accuracy. Results support the use of a combination of logistic regression and ROC curve analysis for setting standards and establishing cut scores with ROC analysis being used to determine the final cut scores to ensure that students experiencing reading difficulties were identified (Silberglitt & Hintze, 2005).

Results of this study support the previous three studies. The performance of over 2,000 students who were administered R-CBM benchmark assessments in the spring of first, second, and third grades and the MCA in the spring of third grade was examined. Like previous

research, results indicated that R-CBM strongly predicted success on the MCA with a moderate to high degree of predictive and concurrent validity as well as a moderate to high degree of diagnostic accuracy. Cut scores were created so that a system of consistent measurement could be determined from the winter of first grade to the spring of third grade. The relationship was strongest for R-CBM assessments that occurred closer in time to the administration of the MCA than those that were further removed in time. R-CBM was able to predict with a high degree of accuracy (greater than 80%) those students who were likely to pass the MCA as far back as the spring of first grade. Students who did not reach the R-CBM cut score in first grade had a better chance of eventually passing the MCA in third grade than students who did not reach the third grade benchmark. Sixty-two percent of students who did not reach the cut-score in the spring of first grade failed the MCA while 68.5 percent of students' not reaching the cut-score in the spring of third grade failed the MCA. Cut scores developed by Silberglitt and Hintze (2005) are consistent with those found by past research; 107 WRC in the spring of third grade, 90 WRC in the spring of second grade, and 49 WRC in the spring of first grade (Silberglitt & Hintze).

Hintze and Silberglitt (2005) sought to compare three statistical and methodological approaches discussed in the previous study (discriminant analysis, logistic regression, and receiver operating characteristic curves) to standard setting and determining cut scores using R-CBM and performance on high-stakes tests. An advantage of discriminant analysis and logistic regression are that they can be used to maximize correct classification. Discriminant analysis tries to maximize true positives, students likely to fail the MCA, and true negatives, students likely to pass the MCA. Logistic regression attempts to maximize only true positives. ROC analysis allows a number of different cut scores to be used across a variety of assessment

situations. Each statistical method will produce different cut scores and have implications for predicting performance on high-stakes assessments (Hintze and Silberglitt, 2005).

The findings of Hintze and Silberglitt (2005) support the use of R-CBM as a powerful predictor of global measures of reading. Consistent with previous studies, results indicate that R-CBM was more strongly correlated with the MCA when the two assessments were collected in closer proximity as compared to farther apart in time. R-CBM measures were also strongly correlated with each other, with those measures collected within a particular grade level more highly correlated than measures across grade levels (Hintze & Silberglitt).

Cut-scores derived using R-CBM successively across grades lead to improved accuracy in identifying students who were likely to fail the MCA. Using R-CBM to set cut scores in a successive manner from one benchmarking period to the next across grades appeared to be a more accurate and efficient method than using high-stakes tests consistently as the criterion regardless of the grade level. In addition to determining risk status for the probability of failing high-stakes tests, established cut-scores were used for identifying students at risk for developing reading problems and students in need of additional reading interventions. Using R-CBM as a school-wide screening measure in the general education setting was supported (Hintze & Silberglitt, 2005).

Each statistical procedure used was able to set cut scores that yielded adequate levels of diagnostic accuracy and efficiency. Each approach resulted in cut scores that yielded higher levels of specificity and negative predictive power as compared to sensitivity and positive predictive power. These findings are consistent with the results of Silberglitt and Hintze (2005). Hintze and Silberglitt recommend using ROC analysis when the goal is to create different cut scores for different types of decisions, such as screening or classification. If one set of scores is

preferred for one set of decisions, discriminant analysis or logistic regression are suitable alternatives.

The Developmental Reading Assessment

The DRA is a literature-based instructional reading and assessment program used to help teachers assess and document student's reading performance over time. The DRA was developed by the Upper Arlington School District to identify students at risk for reading failure (Pearson Learning Group).

During field-testing of the DRA in 1996, teachers described the DRA as being helpful in describing reading behavior. They also stated that it complimented classroom instructional activities. Williams (1999) examined the inter-rater agreement of teachers using the DRA K-3 assessment and the internal consistency of the DRA. A sample of 306 students from kindergarten through third grade was included in this study. Participating teachers were instructed to audiotape the DRA conference. These tapes were then sent to a second and then a third person to rate. Rasch rating scale analyses were conducted on the data. Results revealed that the inter-rater agreement between the original teacher and a second rater was .80, barely adequate for screening assessments. Inter-rater agreement across all three raters was .74 (Williams).

Williams (1999) gathered additional data to help establish the construct validity of the DRA. Individual scores on the DRA for second grade students were correlated with their scores from the fall of third grade on the Iowa Test of Basic Skills Subscales: Vocabulary, Reading Comprehension, and Total Reading. All correlations were significant with the highest and most meaningful correlation found between the DRA and Total Reading (.71) (Williams).

Weber (2000) examined the observer agreement of teachers using the DRA. To examine the consistency between teachers, a group of 10 teachers watched behind a one-way mirror while an expert conducted the DRA conferences with four students. Each teacher scored the accuracy of the students' oral reading independent of the expert and other teachers. The percentage of agreement between a teacher and an expert rater was uniformly high, indicating high rates of scorer validity (Weber). Weber also examined the test-retest reliability of the DRA with three weeks between the first administration and the second administration. The obtained correlation coefficients ranged from .92 to .99 and all were statistically significant. These results indicate that the DRA provides consistent evaluations of a students' independent reading level over time (Weber).

Weber (2000) sought to determine the extent to which students' independent reading level on the DRA is predictive of student performance on the reading comprehension section of the Iowa Tests of Basic Skills. The obtained correlation coefficients range from .54 to .83, all of which are statistically significant. Results indicate that performance on the DRA is predictive of performance on the reading comprehension section of the Iowa Test of Basic Skills. The results suggest a moderate level of criterion validity (Weber).

Buchanan (2002) compared DRA Independent Reading Levels for various groups of students. A statistically significant difference was noted between change in independent reading level and racial group. Caucasian students' independent reading level increased more than that of African American and Hispanic students. The change in DRA level was also found to be significant when compared to Section 504 status and IDEA status (Buchanan).

No known previous researchers have examined the relationship between the DRA and high-stakes tests. When considering how frequently the DRA is used among school districts in

the United States and the high value placed on the results of high-stakes assessments, this is an area that must be addressed in the research.

Summary

In an era of high-stakes assessment, an evaluation system is needed that can be used simultaneously with instruction to identify students at-risk for reading difficulties and prevent long-term reading failure. Formative assessments such as R-CBM and the DRA need to be further evaluated to assess their utility as screening measures for identifying students at risk for reading difficulties and likely to not achieve satisfactory scores on state assessments. If these students can be identified prior to failing a state assessment, they can receive additional reading supports and interventions sooner. The previous studies must be replicated in more diverse settings, over longer periods of time, and with a broader array of high-stakes evaluations.

Chapter III

Method

Participants

One hundred ninety four students from an urban school district in Western New York participated in this study. Participants were enrolled in four schools in the district that are all recipients of a Reading First grant. Third grade students during the 2004-2005 school year who participated in both the DIBELS-ORF (DORF) and DRA assessments and also took the fourth grade ELA in 2006 were involved in the current study. Demographic information specific to the participants was not available. The ethnic makeup of the students in the district is 65% African American, 22% Hispanic, 12% white, and 2% Native American, Asian and other minorities. Thirty-five different languages are spoken within the student population. Based on family income, 88% of the district's students are eligible for free or reduced-price lunch. Seventeen percent of the student population receives special education services.

Confidentiality

A signed statement of confidentiality was generated between the researcher and the school district that provided the data. Identification numbers were used to replace identifying student information before the data was given to the researcher. Furthermore, access to the data was restricted to the primary investigator and the university thesis advisor.

Predictors

Student reading performance on the DORF using reading curriculum based measurement (R-CBM) and the DRA at each benchmarking period served as the predictors in this study.

DIBELS Oral Reading Fluency (DORF). DIBELS Oral Reading Fluency (Dynamic Measurement Group, 2002) is a standardized, individually administered test of accuracy and

fluency with connected text. The DORF is based on the development of Reading Curriculum-Based Measurement by Stanley Deno and his colleagues at the University of Minnesota (Deno, 1985). The Reading First grant program mandates that all students be assessed using the DORF.

A number of studies have established the technical adequacy of R-CBM procedures. Shinn at al. (1992) found strong support for the efficiency of oral reading fluency as a measure of reading proficiency and comprehension. The content, criterion, and construct validities of R-CBM as well as alternate-form and test-retest reliabilities are well documented and sustained (Good, Simmons, & Kame'enui, 2001). In 1989, Martson reviewed existing research on CBM and revealed test-retest reliabilities above .90. Inter-rater agreement was found to be .99 (Martson, 1989). Studies demonstrate strong criterion validity with respect to published norm referenced reading achievement tests and criterion-referenced basal reading mastery tests. Criterion-related validity has been reported to range from .52 to .91, with most above .80 (Fuchs & Fuchs, 1993; Shinn et al., 1992; Martson).

During each benchmarking period (fall, winter, and spring), students in the current study were given three 1-minute timed reading passages. Each reading passage consisted of approximately 250 words. Words omitted, substituted, and hesitations for more than three seconds were scored as errors. The number of words read correct per minute was recorded as the oral reading fluency rate. The median score was used during each benchmarking period. Different standard benchmark reading assessment passages were used during the course of this investigation and were based on the goal level of reading for each grade level. Benchmark scores for DORF are provided through the assessment materials. These scores are an established standard of performance that can be used to indicate a student's oral reading fluency abilities compared to same grade peers. Benchmarks for each assessment period for third grade students

on the DORF are as follows: fall benchmark = 77 WRC, winter benchmark = 92 WRC and the spring benchmark = 100 WRC.

Developmental Reading Assessment (DRA). The Developmental Reading Assessment (DRA; Beaver, 2002) is an individually administered diagnostic instrument. It is designed to determine a child's independent reading level and to assess students' strengths and weaknesses in relation to engagement, oral reading fluency, and comprehension. The DRA was administered to participating students in a one-on-one format between student and teacher (Pearson Learning Group). The district in which this study was conducted mandated that all students in kindergarten through third grade be assessed using the DRA.

The DRA K-3 consists of two assessment texts, representing a range of text difficulty, indicated on a scale from A to 44. Teachers are able to document student reading development over time. A graph is provided for monitoring each student's independent reading level progress (Pearson Learning Group).

According to the manual, the DRA K-3 takes approximately 10-20 minutes to administer (Pearson Learning Group, 2003). Administration procedures vary depending on the students reading level; teachers select reading material for students reading below Level 2 while those reading above Level 2 are instructed to choose from a pre-arranged set of books. Students predict the outcome of the book and are asked to read aloud or silently, depending on their reading level. Next, teachers record oral reading while students read aloud. Those students reading above Level 2 are instructed to retell the story. Teachers then ask students about their reading preferences. Teachers complete an observation guide during the conference. Comprehension scores, based on the observation guide, range from 6 to 24. Each student is assigned an independent reading level. This is based on accuracy, which is the level the student

reads at 94% or greater; fluency, based on teacher judgment and; comprehension, based on scores of 16 or higher on the comprehension rubric (Pearson Learning Group).

The district in the current study utilized its own DRA benchmark goals in which individual student performance can be compared. Benchmark scores for the DRA used by the district that are indicative of an increased likelihood of reading success are as follows: fall benchmark = 28 and spring benchmark = 34. These district benchmarks are based on assessment at a students instructional level; students were assessed with texts that would be appropriate for classroom instruction but would be more difficult than what the child would be expected to read on his or her own. Assessment at a students independent reading level, in contrast, would involve assessing at a reading level in which a child was successfully reading on his or her own.

The manual presents adequate evidence of the DRA as a reliable and valid measure. Results of a study conducted by Weber (2000) indicate that the DRA provides consistent evaluations of a student's independent reading level over time. Williams (1999) found that the agreement between the original teacher and a second rater was good (.80). Weber (2000) found that the percentage of agreement between a teacher and an expert rater was uniformly high, indicating high rates of scorer validity.

Dependent Measure

Performance on state mandated reading assessments was measured by student performance on the fourth grade English Language Arts (ELA) exam and will serve as the criterion (dependent measure) in this study. The ELA is the annual summative high-stakes evaluation used in New York State. Using student performance on the ELA as the criterion standard, risk indicators were developed for predicting which students are not likely to experience success on the ELA.

English Language Arts Exam (ELA). Fourth grade students in New York State participate in a three-day-long standardized English Language Arts (ELA) exam. The students in this study participated in the exam in January 2006 (New York State Education Department).

The New York State Testing Program 2006: English Language Arts, Grades 3 through 8 Technical Report indicates that the Chronbach's Alpha total test reliability coefficient for the fourth grade ELA is .88. Chronbach's Alpha reliability coefficients across various racial, ethnic and gender subgroups for the fourth grade ELA were all greater than .85. The technical report states that the ELA 3-8 maintains a high level of internal consistency, providing evidence of construct validity (New York State Education Department, 2006).

The New York State Education Department (NYSED) has developed four Learning Standards for all students that are incorporated into the ELA exam. Standard one states that students will read, write, listen, and speak for information and understanding. Standard two incorporates language for literacy response and expression, Standard three involves language for critical analysis and evaluation, and Standard four incorporates the use of language for social interaction. The ELA is used to measure the extent to which individual students achieve these learning standards and to determine whether schools, districts, and the state meet the required progress targets specified in the New York State accountability system (New York State Education Department).

The ELA exam is a mixture of multiple-choice, short open-ended and long-open-ended questions. It also contains a reading comprehension part in which the teacher reads aloud a passage on which the children take notes and then write two short answer and one extended answer response. Students earn a performance level score of 1, 2, 3, or 4 on the ELA. Students receiving a score of 1 or 2 are considered not to be meeting grade level expectations in reading

while those that receive a score of 3 or 4 are considered to be meeting grade level reading expectations (New York State Education Department).

Procedures

Data from student performance on DORF and DRA measures as well as performance on the ELA was examined from an existing data base. Data was collected by teachers and reading specialists employed by the school district. Participants were assessed three times with the DORF (fall, winter, and spring) and two times with the DRA (fall and spring) measures beginning in the fall of third grade. Students were administered the ELA in the winter of fourth grade. Administration procedures were provided by the New York State Department of Education.

Data Analysis

Descriptive statistics were calculated for each measure. The correlation between DORF and the ELA and between the DRA and ELA was examined across each benchmarking period. Next, screening risk indicators were established for identifying students at risk for failing the ELA. These cut-scores can be useful for identifying students in need of additional reading supports and interventions. Risk indicators were developed by conducting receiver operating characteristic (ROC) curves and screening outcome matrices.

The ROC curves and screening outcome matrices were created for the DORF in the fall, winter, and spring, and the DRA in the fall and spring. The screening outcome matrices were used to analyze the following data for each possible risk indicator: True positives (TP), which represent the number of students who were correctly identified at-risk based on the screening measure (identified at-risk and failed ELA); false positives (FP), the number students incorrectly identified as at-risk (identified at-risk and passed ELA); true negatives (TN), the number of

students who were correctly identified as not at-risk based on the screening measure (identified not at-risk and passed the ELA); and false negatives (FN), the number of students incorrectly identified as not at risk (identified as not at-risk and failed ELA) (Glover & Albers, 2007; Hintze, Ryan, & Stoner, 2003).

From these matrices, each possible cut score's sensitivity, specificity, positive predictive value, negative predictive value, and correct classification were calculated. The sensitivity index is an indicator of whether the screening instrument correctly identified those students who were later found to be at risk (TP/TP+FN). The specificity index is an indicator of whether the screening instrument correctly identified those students later found to be not at-risk (TN/FP+TN). A trade-off exists between sensitivity and specificity. As sensitivity increases, decreases in specificity are observed and vice versa (Glover & Albers, 2007; Hintze, Ryan, & Stoner, 2003). Positive predictive value (PPV) is an indicator of the proportion of students who were correctly identified as at risk (TP) out of all students who were identified as at risk on the screening instrument (TP+FP). Negative predictive value (NPV) is an indicator of the proportion of students who were correctly identified as not at risk (TN) out of all students identified as not at risk on the screening instrument (FN+TN). Correct classification (CC) is the number of students correctly identified as at risk or not at risk based on the screening measure (TP+TN/total) (Glover & Albers; Hintze, Ryan, & Stoner). Optimal cut scores were chosen based on these criteria. Receiver operating characteristic (ROC) curves were used to plot the sensitivity and specificity of the predictor variables (DORF and DRA) for all possible values of the cut score.

To answer the research question: (a) To what degree does student performance on R-CBM and the DRA screening measures in third grade correlate with fourth grade ELA performance?,

the researcher examined descriptive statistics and calculated Pearson correlation coefficients. To answer the remaining questions: (b) What score on R-CBM in third grade can be used to establish a screening risk indicator that can be used to identify students likely to not achieve satisfactory scores on the fourth grade ELA exam?, and (c) What score on the DRA in third grade can be used to establish a screening risk indicator that can be used to identify students not likely to achieve satisfactory scores on the fourth grade ELA exam?, the researcher conducted receiver operating characteristic (ROC) curves and created screening outcome matrices.

Chapter IV

Results

Descriptive Statistics

Table 1 contains the descriptive statistics for all children on the DORF, DRA, and ELA measures. According to the DIBELS benchmark goals and indicators of risk (www.dibels.uoregon.edu), the data suggests that the average student's DIBELS performance is within the "some risk" status category. According to the DRA performance goals developed and used by the district, the average student's DRA performance is within the "at-risk" status category. Examination of the distribution of scores suggests variability within scores assuming a normal distribution within the DORF-S and DRA-F variables. Examination of the distribution of scores suggests the DORF-F and DORF-W were slightly positively skewed with more students scoring in the lower range of words read correct per minute. DRA-S appears slightly negatively skewed. The ELA appears slightly positively skewed, although the majority of students scored at levels 2 and 3. Variables skewed in the opposite direction will result in lower correlations. Because the ELA score is an ordinal variable and the DORF and DRA scores are continuous variables, correlations will be approximate. Overall, the data indicates the variability of scores was adequate for further analysis.

Table 1

Descriptive Statistics for Total Sample (N = 194)

Variable	Mean	(SD)	Min.	Max.	Skew. Statistic
ORF-F	64.30	(25.51)	11	141	.491
ORF-W	81.66	(31.75)	12	191	.838
ORF-S	98.26	(32.41)	15	185	.030
DRA-F	13.63	(3.08)	5	20	-.357
DRA-S	15.95	(2.66)	7	20	-.740
ELA PL	2.27	(.714)	1	4	-.336

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; ELA PL = English Language Arts Exam Performance Level.

Pearson Correlation

Table 2 contains the Pearson correlation coefficients for all children on the DORF, DRA, and the ELA performance level. The Pearson correlation was used to examine the relationship between DORF measures and the ELA and between DRA measures and the ELA. All correlations were positive and significant at the 0.01 level (2-tailed). Statistically significant correlations were obtained between the administrations of each screening measure during each benchmarking period as well as between the two screening measures. The highest correlation between an independent variable and the ELA existed between the DRA-S in third grade and the ELA in fourth grade, $r\,(191) = .548$, $p < .01$. Correlations between the DORF and ELA and between the DRA and ELA decreased as the amount of time between test administrations increased.

Table 2

Pearson Correlations for Scores on the DORF, DRA, and ELA

Measure	ORF-F	ORF-W	ORF-S	DRA-F	DRA-S	ELA PL
ORF-F	1	.894	.862	.665	.657	.472
ORF-W	.894	1	.850	.670	.625	.479
ORF-S	.862	.850	1	.599	.677	.516
DRA-F	.665	.670	.599	1	.848	.534
DRA-S	.657	.625	.677	.848	1	.548

Note. All correlations are significant at the .01 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; ELA PL = English Language Arts Exam Performance Level.

Diagnostic Accuracy Analysis

Receiver operating characteristic curves. In an effort to explore the relationship between DORF and the DRA with the ELA, a series of receiver operating characteristic (ROC) curves were developed. Pairs of values were plotted, with (1-specificity) on the X axis and sensitivity on the Y-axis, yielding the curves in Figures 1-3. These figures represent the DORF and DRA at each of the benchmarking periods modeled against the performance level on the ELA. They provide a graphical representation of the trade-off between sensitivity and specificity among the range of all possible cut points for each screening measure (Streiner & Cairney, 2007). The more the ROC curve deviates from the dotted line and tends toward the upper left corner, the better the sensitivity and specificity of the assessment measure. The optimal cut point that is closest to the upper left corner is the one that minimizes the overall number of errors (Streiner & Cairney). Assessment instruments that do not discriminate well have curves that are nearer the reference line. The reference line indicates the relationship between true-positive and false-positive rates when an assessment instrument yields no useful diagnostic information beyond chance (Hintze, Ryan, & Stoner, 2003).

The area under the curve (AUC) represents a measure of test accuracy. The AUC represents the probability that the screening measure will correctly identify a student as at risk for failing the ELA. Screening measures with a larger AUC possess greater discriminatory abilities. The AUC can be compared to the null hypothesis, that the test has no predictive value, represented by an AUC of .50. An AUC between .50 and .70 is low; between .70 and .90 is moderate; and above .90 is high. These values allow the comparison between the accuracy of different screening measures. The measure with the higher AUC is preferable (Tape, 2007; Streiner & Cairney, 2007). When establishing the validity of screening measures, such as the

DORF and DRA, the asymptotic significance, or p-value, should also be examined. The asymptotic significance represents the significance of the AUC. A p-value less then .05 suggests that the null hypothesis should be rejected and that the screening measure significantly predicts failure on the ELA at a rate greater than chance alone.

Figure 1 represents the ROC curve for the DORF-F and DRA-F measures in predicting failure on the fourth grade ELA exam. Based on the figure, it can be seen that the DORF-F measure appears to correctly identify more students who later failed the ELA and are therefore in need of reading supports and interventions. Although moderate to high levels of sensitivity are observed using risk indicators in the range of 60 to 80 for DORF-F and 14 to 18 for DRA-F, moderate to weak levels of specificity are noted. DORF-F has an AUC of .642 while the DRA has an AUC of .595, indicating that DORF-F is the preferred performance measure because it more accurately identifies students who are at risk for reading difficulties. The DORF-F has an asymptotic significance of .002 and the DRA-F has an asymptotic significance of .037. Both are less then .05, suggesting that the DORF-F and DRA-F predict failure on the ELA significantly better than chance.

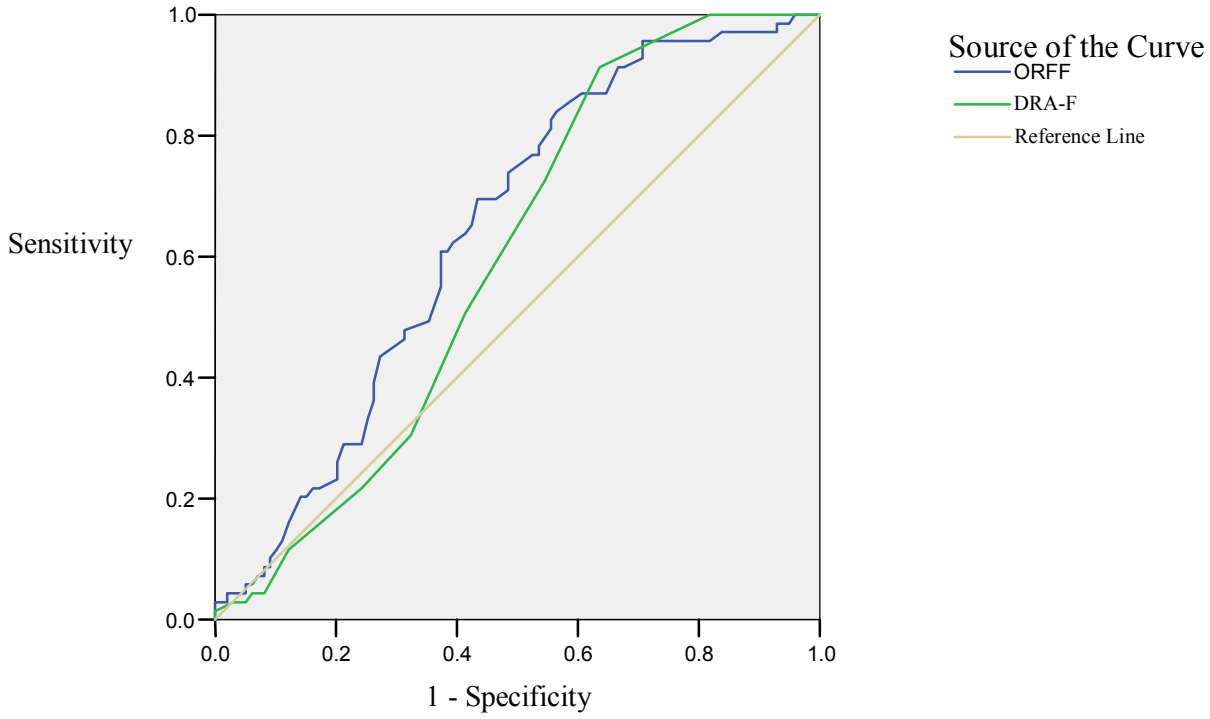
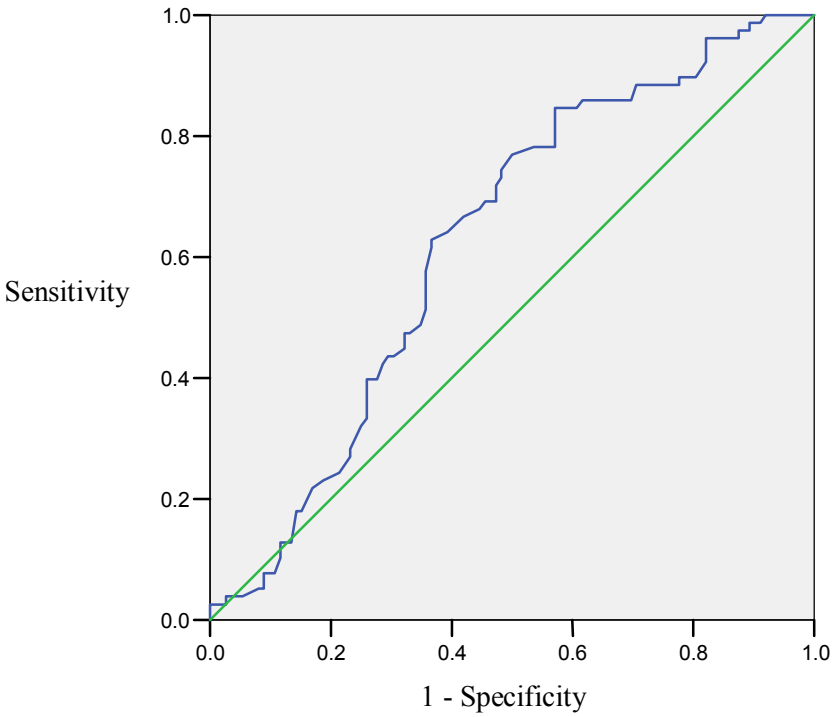


Figure Caption

Figure 1. ROC curves for third grade fall benchmarking measures predicting failure on the fourth grade ELA exam

Figure 2 represents the ROC curve for the DORF-W measure in predicting failure on the fourth grade ELA. This screening measure appears to be moderately effective in identifying which students are in need of additional reading supports and intervention. It appears to over-identify students as at-risk and in need of reading supports and interventions. The DORF-W screening measure has an AUC of .627, in the low range for accurately predicting performance on the fourth grade ELA exam. The DORF-W screening measure has an asymptotic significance value of .003, indicating that its ability to predict failure on the ELA is significantly better than chance.



.

Figure Caption

Figure 2. ROC curve for third grade winter benchmarking measure predicting failure on the fourth grade ELA

Figure 3 provides the results of the ROC curve analysis for the DORF-S and DRA-S measures used to predict failure on the fourth grade ELA exam. Although both screening measures exhibit moderate diagnostic accuracy, the ORF-S screening measure most successfully balanced the sensitivity and specificity indexes to effectively identify students who went on to fail the ELA. The DRA-S screening measure appears to over identify more students as at-risk who later experienced success on the ELA. The DORF-S measure had an AUC of .674, the largest among the screening measures in this study, and the DRA-S had an AUC of .625. Both values are in the low range with respect to their ability to accurately predict performance on the fourth grade ELA. The DORF-F had an asymptotic significance value of .000 while the DRA had an asymptotic significance value of .004. Both measures predict failure on the ELA significantly better than chance alone.

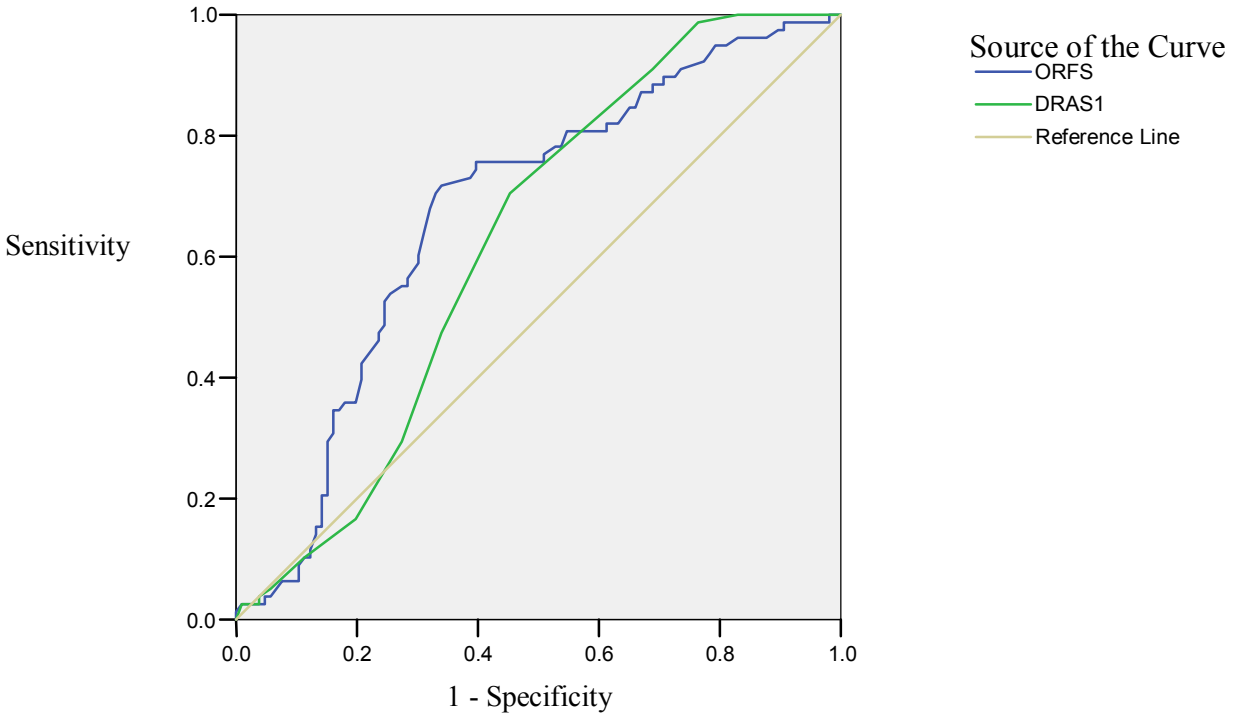


Figure Caption

Figure 3. ROC curve for the third grade spring benchmarking measures predicting failure on the fourth grade ELA

Diagnostic accuracy statistics. A series of analyses were completed using the DORF and DRA as predictor variables and the ELA as the criterion measure. These analyses examined the diagnostic accuracy of the DORF and DRA as screening measures for identifying which students need additional reading supports and interventions in order to experience success on the ELA. In addition, risk indicators were developed for each benchmarking period. A screening instrument outcome matrix, the 2x2 table located in Table 3, was created for each predictor variable (Glover & Albers, 2007; Hintze, Ryan, & Stoner, 2003).

Table 3

Sample 2x2 Screening Instrument Outcome Matrix

Screening Assessment Outcomes		
Criterion measure	Screening identification	
Performance	At risk	Not at risk
Poor outcome	True Positive (TP)	False Negative (FN)
Adequate Outcome	False Positive (FP)	True Negative (TN)

The information above was summarized by means of sensitivity, specificity, positive predictive value, negative predictive value, and correct classification for each screening measure. A screening outcome matrix, similar to that in Table 3, was created for selected cut scores representing a range of possible performance outcomes for each assessment measure. From these matrices, each cut score's sensitivity, specificity, PPV, NPV, and CC was calculated. The results of these studies are contained in Table 4. Based on this information, a single cut-score was chosen for each screening measure at each benchmarking period to predict students not likely to achieve satisfactory scores on the ELA. These scores can be used as screening risk indicators for determining when a child needs interventions and additional reading supports. Chosen risk indicators selected for each measure are marked with an asterisk to distinguish them in Table 4.

Table 4

Performance of the DORF and DRA over a Range of Cut-Scores using the ELA Performance Level Scores as the Criteria

ELA					
ORF F cut score	Sensitivity	Specificity	PPV	NPV	CC
50	.382	.850	.777	.500	59%
60	.642	.750	.733	.612	68%
70*	.736	.650	.736	.650	70%
75	.830	.563	.715	.714	72%
80	.858	.488	.689	.722	70%

ELA					
ORF W cut score	Sensitivity	Specificity	PPV	NPV	CC
60	.375	.904	.830	.536	61%
70	.548	.807	.781	.588	66%
80*	.712	.699	.748	.659	71%
90	.846	.578	.715	.750	73%
100	.914	.434	.669	.800	70%

ELA					
ORF S cut score	Sensitivity	Specificity	PPV	NPV	CC
80	.407	.915	.863	.540	63%
90	.593	.830	.821	.607	69%
100*	.769	.720	.783	.702	75%

110	.824	.585	.724	.716	72%
120	.880	.439	.674	.735	69%

ELA

DRA F cut score	Sensitivity	Specificity	PPV	NPV	CC
10	.200	.987	.950	.490	54%
12	.410	.811	.736	.517	59%
14*	.768	.568	.695	.652	68%
16	.979	.324	.650	.923	69%
18	1.000	.095	.602	1.000	64%

ELA

DRA S cut score	Sensitivity	Specificity	PPV	NPV	CC
12	.179	.987	.950	.476	53%
14	.415	.900	.846	.537	62%
16*	.759	.675	.772	.659	76%
18	.981	.313	.654	.926	69%

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; ELA PL = English Language Arts Exam Performance Level.

* Denotes the cut-score chosen to most effectively predict failure on the ELA

Optimal cut-scores were chosen by considering each measures sensitivity, specificity, PPV, NPV, and CC across difference performance levels. The goal was to choose a cut-score with high enough sensitivity so that students at-risk will be identified and can receive the supports they need. Given that the purpose of the study was to examine the utility of DORF and DRA as screening measures used to identify students in need of reading supports and interventions, a higher sensitivity index is warranted. The CC index was also taken into consideration when making the decision. The PPV, the proportion of students correctly identified as at-risk out of all students identified as at-risk was also taken into consideration. The researcher sought to chose a cut-score with high PPV to limit the chance that the measure is over-identifying students as at-risk.

Cut-scores of 70 or fewer WRC on the ORF-F measure, 80 or fewer WRC on the ORF-W measure, and 100 or fewer WRC on the ORF-S measure in third grade were chosen as the risk indicators that best predicted student failure on the fourth grade ELA. Risk indicators of 14 or less on the DRA-F measure and 16 or less on the DRA-S measure in third grade were chosen as the scores that best predicted student failure on the fourth grade ELA.

Ideally, levels of sensitivity and specificity are generally considered adequate at approximately .75 or higher (Glover & Albers, 2007). As can be examined in Table 3, chosen cut scores in the current study have sensitivity indexes that range from .712 to .769 for the DORF, .759 to .768 for the DRA and specificity indexes that range from .650 to .720 for the DORF and .568 to .675 for the DRA (Hintze, Ryan & Stoner, 2003). The optimal cut-scores chosen for the screening measures in the current study correctly identified a larger proportion of students who later failed the ELA exam (sensitivity) than correctly identified the proportion of students who later passed the ELA (specificity).

Table 5 and Table 6 can be used to examine how the risk indicators established in the current study compare to the pre-established DORF and DRA critical performance levels currently used by the district. DORF benchmark critical performance levels endorsed by DIBELS and used in the district were established by Good et al. (2001). The critical performance levels established in the current study for identifying students in need of additional reading supports and interventions coincide with those established by Good et al. (2001). The screening risk indicators developed in this study for identifying when a child is in need of additional reading supports and interventions all fall within the "some risk" range developed by Good et al. (2001).

Table 5

Third grade DORF Benchmark Critical Performance Levels

Benchmark	A-risk	S-risk	L-risk
Fall DORF	<53	53-76	>77
Winter DORF	<67	67-91	>92
Spring DORF	<80	80-109	>110

Note. A-risk = At-risk; S-risk = Some-risk; L-risk = Low-risk.

The district in which this study was conducted used the DRA critical performance levels found in Table 6 for identifying when students are in need of additional reading interventions and supports. These cut-scores were sent to the researcher by the district but it is unclear how and by whom they were developed. According to the critical performance levels in Table 6, every student in the current study would have been identified as at-risk and in need of reading support. Therefore, the scores identified in the current study appear to be a more accurate approach for screening children and identifying students in need of additional supports and interventions.

Table 6

Third grade DRA Benchmark Critical Performance Levels

Benchmark	A-risk	S-risk	L-risk
Fall DRA	<20	24	28
Spring DRA	<28	28-30	34

Note: A-risk = At-risk; S-risk = Some-risk; L-risk = Low-risk.

Chapter V

Discussion

The purpose of the present study was to examine how formative assessments such as the DORF and DRA can be used as screening measures for identifying students who are at risk for reading difficulties and are not likely to achieve a satisfactory score on the state reading assessment. If a screening system can be implemented for identifying these students before the state assessment, they will be able to receive reading supports and services sooner. Risk indicators were created for aiding in the identification of students not likely to achieve satisfactory scores on the state reading assessment. These cut scores can also be used as a tool for monitoring student progress and can aid in setting goals for students once they begin receiving interventions.

Relationship of the DORF and DRA to the ELA

The current study sought to examine the degree to which student performance on R-CBM and DRA screening measures in third grade correlated with fourth grade ELA performance. Results indicate that reading performance on the third grade DORF and the DRA were found to significantly correlate with fourth grade ELA performance for all benchmarking periods. Student performance on the DORF in the fall, winter, and spring of third grade, and the DRA in the fall and spring of third grade, all significantly correlated with student performance on the fourth grade ELA exam. The DRA-S measure appeared to have the strongest correlation with the fourth grade ELA. This is not surprising considering the time of year that both measures are administered.

Squared correlations indicate that the DORF, DRA, and ELA share approximately 22 to 30 percent of the variance with the ELA. This suggests that the DORF and DRA may measure

one of several constructs assessed by the ELA. The majority of the variance may be accounted for by other factors, such as a student's ability to write, listen, and speak for information and understanding and a student's ability to use language for critical analysis, evaluation, and social interaction. This is a possible reason why the diagnostic accuracy of the DORF and DRA was not higher.

Risk Indicators and Diagnostic Accuracy

The study also sought to establish screening risk indicators for R-CBM and the DRA in third grade to aid in the identification of students likely to not achieve satisfactory scores on the fourth grade ELA exam. Results suggest that with a moderate amount of diagnostic accuracy, the DORF and DRA can be used as screening measures for identifying third grade students in need of reading supports and interventions. Risk indicators were created for each measure during each assessment period for predicting which students are at risk for failing the fourth grade ELA exam.

Data provided through the ROC curve analysis suggests that DORF and DRA measures for all benchmarking periods were able to predict failure on the fourth grade ELA significantly better than chance. Both measures were able to identify students at risk for reading difficulties with a low to moderate degree of diagnostic accuracy. The AUC for all measures ranged from .595 for the DRA-fall to .674 for the DORF-spring. All screening measures had asymptotic significance levels below .05, supporting the DORF and DRA as valid measurements for predicting performance on the fourth grade ELA.

When evaluating the adequacy of the DORF and the DRA as screening instruments, the practical implications associated with indices of predictive validity were considered. Positive predictive value and sensitivity are often considered most important when determining a

screening instrument's adequacy. When screening measures have a lower positive predictive value, they have a greater chance of over-identifying students as at-risk. Serving incorrectly identified students may result in reduced opportunities for learning or growth, overuse of programming and school resources, and increased stress among school personnel (Glover & Albers, 2007). Sensitivity is also important to consider. A low sensitivity value may indicate that the measure is under-identifying students who are at risk. As a result, these under-identified students may not receive the services and supports that they need to succeed academically. The challenge is to set a cut-score that maximizes both sensitivity and specificity to its fullest potential (Glover & Albers). Because the current study sought to use screening measures to identify children in need of reading supports and interventions, the researcher tolerated more false positives, identification of students at risk who later passed the ELA than false negatives, identification of students not at risk who later failed the ELA. (Hintze, Ryan, & Stoner, 2003)

For both the DORF and DRA measures, the suggested risk indicators result in higher levels of sensitivity (.73 to .76 for DORF and .75 to .76 for DRA) than specificity (.65 to .71 for DORF and .56 to .67 for DRA). Sensitivity for both measures is generally considered adequate, while specificity is moderate. The risk indicators created in the current study may lead to higher percentages of false positives; these cut scores may identify a higher percentage of third grade students as at-risk who later go on to experience success on the fourth grade ELA exam. Because the purpose of this study was to examine the utility of these measures as screening assessments, this is regarded as acceptable. Once a student is identified as at-risk based on these screening measures, it is recommended that follow up assessments be given to further examine difficulties. This can be done in the form of progress monitoring or additional assessments with more comprehensive instruments. Follow-up assessments can serve dual advantages; teachers

can be reassured that students who were incorrectly identified as at risk will be re-examined, reducing the likelihood that supports will be given when they are not needed, and a more thorough assessment can help assist in the planning of specific and targeted reading interventions and supports specific to each individual student's needs. These interventions and supports can assist third grade students in gaining the skills they need in order to experience success on the fourth grade ELA. This helps to balance the risk of incorrect classification.

Implications for Theory

Early identification and intervention are critical to improving reading outcomes. Implementing a screening assessment for predicting future performance on the ELA fits into the problem solving model and a prevention oriented assessment framework which focuses on early prevention of reading failure. Using risk indicators for identifying students who need reading supports and interventions can help eliminate early reading difficulties from becoming later reading failure. Screening risk indicators can be used to progress monitor students found to be at risk based on the universal screening assessment. Progress monitoring can be used to aid teachers in informing instruction and implementing interventions based on student progress.

Risk indicators developed in the current study are fairly consistent with those developed by past research (Good et al., 2001). The current third grade cut scores of 70 WRC or fewer in the fall, 80 WRC or fewer in the winter, and 100 WRC or fewer in the spring are all within the some-risk classification range of benchmark goals developed by Good et al. (2005). The risk indicators developed for the DRA of 14 or fewer in the fall of third grade and 16 or fewer in the winter of third grade can also be used to aide in the identification of students at risk for reading difficulties. These risk indicators add valuable information to the current available research on the DRA. Student's performing below the cut score on either of these screening measures can be

identified and teachers can initiate a problem solving analysis in order to tailor instruction and interventions to the student's individualized reading needs. Once children are identified through a benchmark assessment that uses these risk indicators, progress monitoring can be used to identify students making inadequate progress and further problem solving can take place.

Implications for Practice

The ROC curve analysis provided evidence that the current DIBELS benchmarks created by Good et al. (2001) can be used effectively during screening assessments with this urban population for identifying which students are in need of additional reading supports and which students are likely to experience failure on the ELA. If being used for classification purposes, the potential for false positives is high. Therefore, it is recommended to follow the screening measure up with additional, more comprehensive instruments.

The DRA benchmarks used by the district in this study, according to the ROC analysis, appear inappropriate. They identified every student in the current study as at-risk for failing the ELA. The district benchmarks are much higher than what is actually necessary for a student to be considered not at risk for failure on the ELA. The reason for the discrepancy between the risk indictors created in the current study and those that were used by the district at the time of data collection may be because the data collected in the present study was based on independent reading level, or the level at which the student is successfully reading on his or her own. However, DRA benchmarks at the time of this study were based on students' instructional reading level. Benchmarks geared at instructional reading levels, or the level at which instruction is geared toward, are higher than those geared towards a student's independent reading level.

Practicing school professionals can use the cut scores created in the current study to aide in the identification of students not likely to experience success on the fourth grade ELA. The chosen cut-scores have correctly identified those later found to be at-risk more accurately at the expense of correctly identifying those later found to be not at risk. Based on these risk indicators, more students will be incorrectly identified as at risk than incorrectly identified as not at risk. This is considered acceptable because as a screening measure, it is desired to attempt to identify all students that are at-risk for experiencing reading difficulties and failing the ELA.

Limitations

Limitations of the present study include the fact that the study is based on archival data. This researcher was not directly involved in the process of data collection. The amount of training data collectors received and the fidelity of administration are not known. Because archival data was used, the researcher did not have access to the demographic information specific to the sample or information regarding which students were provided with reading interventions throughout the 2004-2005 and 2005-2006 school years. The impact of having students receive interventions throughout the study is not fully known and may have influenced the risk indicators developed in the current study. If students were receiving interventions throughout the study, the predictive validity of the DORF and DRA may have been diminished.

Because this study is based on an urban population, it is unknown how well the results will generalize to non-urban school populations. A large percentage (53) of students that participated in this investigation did not experience success on the fourth grade ELA exam. This also needs to be considered when discussing how well the results of this study may generalize to other populations. The present results may have more applicability to settings with similar base rates for failure of the ELA.

Directions for Future Research

Further analysis is warranted with a focus on developing DORF and DRA risk indicators for other grade levels, particularly early elementary school. It is imperative that students at risk for developing reading difficulties and not achieving satisfactory scores on the ELA or other statewide assessments be identified at an early age. Young children who are not on the track to becoming proficient readers can therefore receive compensatory reading services at a young age. Exploring different cut scores for different diagnostic purposes is also warranted. If schools are using the DORF or DRA for classification purposes, different risk indicators may be desired. It may also be beneficial to examine how the DORF and DRA can be used to predict different outcomes, not just failure on the fourth grade ELA. Examining how and if these risk indicators correlate with state wide assessments in other states may also be beneficial.

In summary, educators can easily implement a formative assessment system using risk indicators for R-CBM and the DRA to aide in the identification of students at risk for reading difficulties. Implementing such a system fits into a prevention oriented framework of assessment and intervention. If a universal screening system is in place early, students with reading difficulties can be detected before they experience failure on high stakes assessments. Once identified, reading supports and interventions can be provided. A consistent set of cut scores for DORF and the DRA allows for regular, frequent, and valid measurement to reading progress toward a common goal, such as performance on state assessments.

References

Ardoin, S., Witt, J., Suldo, S., Connell, J., Koenig, J., Resetar, J., Slider, N., & Williams, K. (2004). Examining the incremental benefits of administering a maze and three verses one curriculum-based measurement reading probes when conducting universal screening. [Electronic version]. *School Psychology Review, 33*, 218-233.

Allington, R. (2005). NCLB, Reading First, and wither the future? *Reading Today, 33*, 18-23. Retrieved October 27, 2006, from http://web.ebscohost.com.

Beaver, J., & Carter, M. (2003). *Teacher Guide: Developmental Reading Assessment, Grades 4–8.* Parsippany, NJ: Pearson Education, Inc.

Buchanan, T. K. (2002). *Developmental Reading Assessment: Student achievement*. Study conducted for the Louisiana Department of Education, Division of School Standards, Accountability, and Assistance.

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

Deno, S. (1985). Curriculum-based Measurement: The Emerging Alternative. *Exceptional Children*, 52, 219-232.

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

Fuchs, L. & Fuchs, D. (1993). Formative evaluation of academic progress: How much growth can we expect. *School Psychology Review, 22*, 27-49. Retrieved October 26, 2006, from http://web.ebscohost.com.

Fuchs L., & Fuchs, D. (1999). Monitoring student progress toward the development of

reading competence: A review of three forms of classroom-based assessment. [Electronic version]. *School Psychology Review, 28*, 659-671.

Fuchs, L., & Fuchs, D. (2001). What is scientifically-based research on progress monitoring. *National Center on Student Progress Monitoring.* Retrieved October 6, 2006 from www.studentprogress.org

Fuchs, L, Fuchs, D., & Compton, D. (2004). Monitoring early reading development in first grade: word identification fluency versus nonsense word fluency. [Electronic version]. *Exceptional Children*, 71, 7-21.

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 in Reading*, 5, 239-256. Retrieved September 29, 2006, from http://web.ebscohost.com.

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

Good, R., Kaminski, R., Simmons, D., & Kame'enui E. (2001). Using Dynamic Indicators of Basic Literacy Skills (DIBELS) in an outcomes-driven model: Steps to reading outcomes. [Electronic version]. *OSSC Bulletin, 44*, 3-26.

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

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. [Electronic version]. *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. [Electronic version]. *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. [Electronic version]. *School Psychology Review, 34*, 372-386.

Jenkins, J., Fuchs, L., Van Den Broek, P., Espin, C., & Deno, S. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, 95, 719-729. Retrieved September 29, 2006, from http://web.ebscohost.com.

Kauerz, K. (2002). No Child Left Behind policy brief. Retrieved October 31, 2006 from http://ecs.org.

Kim-Sung, K. (2006). Determining adequate yearly progress from kindergarten through grade 6 with curriculum-based measurement. Retrieved October 6, 2006 from www.studentprogress.org.

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.

Lyon, G. (1998). The NICHD research program in reading development, reading disorders and reading instruction: A summary of research findings. *National Institute of Child Health & Human Development, National Institute of Health.*

Madeline, A., & Wheldall, K. (2004). Curriculum-based measurement of reading: recent advances. [Electronic version]. *International Journal of Disability, Development and Education, 51*, 57-82.

Martson, D. (1989). Curriculum-based measurement: What is it and why do it? In M. R. Shinn (Ed.), Curriculum-based measurement: Assessing special children (pp. 19-78). New York: 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. Retrieved October 27, 2006, from http://web.ebscohost.com.

National Reading Panel. (2000). Teaching children to read: An evidenced-based assessment of the scientific research literature on reading and its implications for reading instruction. *Report of the National Reading Panel* [On-line]. Available: hrrp://www.nationalreadingpanel.org/Documents.

Neill, M. (2006). *Preparing teachers to beat the agonies of NCLB*. Retrieved October 6, 2006 from www.eddigest.com.

New York State Education Department. English Language Arts. Retrieved January 8, 2007, from http://www.emsc.nysed.gov/ciai/ela.html.

New York State Educational Department. (2006). New York State Testing Program 2006: English Language Arts, Grades 3-8 Technical Report. Retrieved March 7, 2008 from http://www.emsc.nysed.gov/osa/pub/gr3-8ela06report.pdf.

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

Shinn, M., Good R., Knutson, N., Tilly, D., & Collins, V. (1992). Curriculum-based

measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*, 459-480. Retrieved October 16, 2007, from http://web.ebscohost.com.

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. [Electronic version]. *Journal of Psychoeducational Assessment, 23*, 304-325.

Sopko, Kim. (2002). *Reading First programs: An overview.* Retrieved September 15, 2006 from http://www.nasdse.org/forum/htm

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

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

Tape, T. (2007). *Interpreting Diagnostic Tests*. Retrieved December 3, 2007, from: www.http//gim.unmc.edu/dxtests/ROC3.htm

U.S Department of Education (2002). *Reading First frequently asked questions*. Retrieved September 15, 2006 from http://www.ed.gov/offices/OESE/readingfirst.

VanDerHeyden, A., & Witt, J. (2005). Quantifying context in assessment: Capturing the effect

of base rates on teacher referral and a problem-solving model of identification. [Electronic Version]. *School Psychology Review, 34*, 161-183.

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

Weber, W. (2000). *Developmental Reading Assessment and evaluación del desarrollo de la lectura: A validation Study*. Research paper published by Pearson Learning Group, Parsippany, NJ. Retrieved October 28, 2006, from http://www.pearsonlearning.com.

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

Williams, E. (1999). Developmental Reading Assessment: Reliability study 1999. Unpublished manuscript. Retrieved October 28, 2006, from http://www.pearsonlearning.com/correlation/rsp/*DRA*.doc.