Using Universal Screening Measures to Accurately Predict Middle School Students' Reading Skill Status on a High-stakes State Test: An Investigation of the Psychometric Properties of Vocabulary, Comprehension, and Fluency Measures

Kari J. Terjeson

A dissertation

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Roxanne Hudson, Chair

Elizabeth West

Heather Hebard

Elizabeth Sanders

Program Authorized to Offer Degree:

College of Education

University of Washington

Abstract

Using Universal Screening Measures to Accurately Predict Middle School Students' Reading Skill Status on a High-stakes State Test: An Investigation of the Psychometric Properties of Vocabulary, Comprehension, and Fluency Measures

Kari J. Terjeson

Chair of the Supervisory Committee:

Professor Roxanne Hudson

College of Education: Special Education

Multiple studies have evaluated universal screening measures used in response to intervention (RtI) frameworks to identify early elementary school students at-risk of academic failure (e.g. Catts et al., 2015; Compton et al., 2006; Johnson et al., 2009; Schatschneider et al., 2004). Less research has been conducted with middle school students (Baker et al., 2015; Espine et al., 2010; Decker et al., 2014; Allison & Johnson, 2011). The purpose of this study was: first, to extend the growing body of research to middle school students and evaluate the effectiveness of individual universal screening tools specifically for students in sixth grade. Second, to determine if a combination of screening measures might increase the classification accuracy rate of these tools. While we expected prior year high-stakes assessment to serve as a strong predictor of future performance, we also hoped to explore alternate means of identification when prior year assessment data is not available. It was predicted that a combination of screeners would reduce the number of false negative results and therein improve consequential validity. This study

utilized correlational, logistic regression, predicted probabilities and ROC curve analysis toward this end. Findings of this study provide similar psychometric results reported from prior research (Denton et al., 2011; Jenkins et al., 2007; Kim et al., 2015). Specifically, this study indicated that proximal measures of reading comprehension, such as prior year MSP and the Gates-MacGinitie (Vocab + Comp), outperform distal measures, like ORF and TOSRE (Vocab + Comp), for students in sixth grade. In addition, when prior year assessment or lengthy measures like the Gates-MacGinitie (Vocab + Comp) are not available, specific combinations of reading measures can improve classification accuracy.

Dedication

Samantha, Allie and Erik,

Collectively, you have given me unconditional love; individually, you have taught me how to live without limits. Without you, this experience would not have been possible. You each inspire me every single day.

Thank you for:

…approaching my mistakes with kindness

…understanding that life is an adventure (yes, even when we get lost)

…defending and protecting each other like a pack of wolves

…tolerating my often overwhelming emotions

…supporting me throughout this crazy dream

…sharing your lives with me daily

…knowing that an exhaustive list here would be a dissertation, in and of itself

Remember, your life is not fixed in time or space. Every day is a decision not a race. My greatest wish for each of you is that you marinate in happiness in whatever form that takes. I am proud and grateful beyond expression. I love you all madly!

Love,

Mom

Acknowledgements

Roxanne Hudson and Elizabeth Sanders your guidance and support allowed me to follow my dreams. Thank you for sharing your expertise and encouragement. I am forever grateful.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Statement of the Problem

The No Child Left Behind (NCLB) act of 2001 requires all students to be proficient readers by the end of the 2013-2014 academic school year. The National Assessment of Educational Progress (NAEP) national longitudinal data trends seem to predict our impending failure to comply with NCLB requirements. In 2002, 25% of eighth grade students were at or below basic and could not be categorized as proficient. Nearly a decade later, in 2011, this number was reduced by 1%, to 24%. Although this is a statistically significant reduction, in the context of NCLB and important educational outcomes, it clearly fails to meet expectations. Students with specific learning disabilities (SLD) face an even greater challenge. An alarming 21% of secondary students with learning disabilities are estimated to be five or more grade levels below in reading (Wagner, Newman, Cameto, Garza & Levine, 2005). It is unlikely that students with this type of deficit will become proficient readers.

Concern regarding illiteracy is well founded. Without intervention, students who experience reading problems in early elementary school continue to display significant deficits as they advance to middle school and into adulthood (Juel, 1988; Satz, Fletcher, Clark and Morris, 1981; Scarborough, 1998; Wagner, 2000). Illiteracy not only impacts long term school performance (Deno, 1989) but also is associated with high school dropout (Juel, 1996; U.S. Dept. of Education, 2007), social problems (Kamhi & Catts, 2012) depression (Ofiesh & Mather, 2013), incarceration (Shippen, Houchins, Crites, Derzis, & Patterson, 2010; Snyder & Sickmund, 1995, 2006), emergency room visits, hospitalizations, heart failure and death (Evangelista et al., 2010; Kutner, Greenburg, Jin, & Paulsen, 2006; Shanahan T. & Shanahan C., 2008), and homelessness (Lee , Tyler & Wright, 2010; McGill-Franzen, 1987; Wagner, 2000). Financial

ramifications of illiteracy are well documented. Nationwide, 44% of people with the lowest literacy levels live in poverty, contrast this to a slight six percent of those in the higher literacy levels (Wagner et al., 2005). Using another measure of poverty, unemployment, researchers found nearly 84% of unemployed fathers and 82% of unemployed mothers lacked a high school diploma (Kutner, Greenberg, Jin, Boyle, Hsu & Dunleavy, 2007). Dropout rates for students with learning disabilities are estimated at 31.6% as compared to 9.4% for students with no disabilities (Kutner et al., 2007). Only 11% of students with learning disabilities, as compared to 53% of general education students have attended a four-year post-secondary education within two years of leaving high school (Wagner et al., 2005). The call for action could not be more compelling. Students who fail to gain adequate literacy skill, regardless of the impetus, suffer long term significant reduction in quality of life.

Research demonstrates that early identification and intervention can alter the learning trajectory of students (e.g., Foorman, Francis, Fletcher, Schatschneider & Mehta, 1998; Simmons et al., 2008; Torgesen, 1997, 1999; Wanzek & Vaughn, 2007) and reduce the number of students who ultimately qualify for special education under the specific learning disability (SLD) category (Torgesen, 2000). As students move into adolescence, interventions are often less successful (Wanzek et al., 2013). Prevention science models, such as Response to Intervention (RtI), operate off this premise. Instead of waiting for a teacher recommendation based on failure, RtI relies on early and frequent universal screening of all students to identify those who may need additional help.

Response to Intervention (RtI) is a multi-tiered framework designed to ensure quality core instruction, monitor student progress, and provide appropriate intervention when necessary (Johnson, Mellard, Fuchs, & McKnight, 2006). In addition, recent changes to the Individuals

with Disabilities Education Improvement Act (IDEIA, 2004) allow school districts to use scientific, research-based interventions, such as RtI, as an alternative to the IQ-achievement discrepancy formula for identifying students with SLD. In fact, in some states, including Colorado, Florida, and Illinois, navigating the RtI process has become the only way to identify students for special education services under the category of SLD (National Center on Response to Intervention (NCRTI), 2010).

After ten years in practice, RtI research continues to test essential constructs. While some features of RtI are widely agreed upon: universal screening, multiple tiers of support, early evidence-based interventions, data-based decisions and special education placement (Gersten, Beckman, Clarke, Foegen, Marsh, Star & Witzel, 2009; Haager, Klingner, & Vaughn, 2007; Johnson, et al., 2006); many specific guidelines are still in question. Research regarding highly predictive universal screening, progress monitoring, decision rules, intensity of interventions and special education placement does not provide specific answers (Jenkins, Schiller, Blakorby, Thayer & Tilly, 2013; O'Connor & Klingner, 2010). In addition, as Pyle and Vaughn (2012) point out, "The research on Response to Intervention (RtI) with secondary students is scant… (p.1)" At first glance, it may seem that RtI should be limited to elementary students since it is an early identification construct by definition. However, as Leach, Scarborough & Rescorla (2003) state, there seems to be a second wave of students whose initial markers of reading difficulty begin beyond third grade. They argue that this group of students is not merely late identified, but late emergent. With this group of students in mind, there is a need for research on universal screening at the secondary level.

The goal in an RtI system is to identify precursors to a problem before an academic or behavioral problem becomes intractable. Intervention in an RtI model does not, at least initially,

translate into special education placement. This proactive approach to identification and intervention is arguably the most important distinguishing factor of RtI. In order for RtI to be successful, however, universal screening tools must accurately identify students in need of instruction beyond the core program. The National Center on Response to Intervention (2010) describes universal screening as

> Screening conducted to identify or predict students who may be at risk for poor learning outcomes. Universal screening tests are typically brief, conducted with all students at a grade level, and followed by additional testing or short-term progress monitoring to corroborate students' risk status. (Universal Screening Section, para.1).

The underlying assumption in this definition is that the assessment will, in fact, identify the correct students in need of intervention to begin with. This assumption can be tested using classification accuracy.

Adopted from the medical model in research, classification accuracy is a rigorous method of evaluating universal screening tool effectiveness. The procedure measures the number of false negatives (students who fall above benchmark on the screening measure yet fail the outcome measure) and false positives (students who fall below benchmark on the screening measure but still pass the outcome measure) and compare these "misdiagnoses" with true negatives (students who fall above benchmark on the screening measure and pass the outcome measure) and true positives (students who fall below benchmark on the screening measure and fail the outcome measure). These measurements are described in terms of sensitivity and specificity. Researchers have called for 90% sensitivity levels on screening tools (Jenkins, Hudson, & Johnson, 2007);

that is to say that the screening tool will identify 90% of the students who will not pass the outcome measure.

Universal screening measures should be quick to administer, simple to score and lead to valid inferences about projected student outcomes (Hosp & Ardoin, 2008). Oral reading fluency (ORF) is arguably the most common universal screening measure used in RtI models beyond second grade. In a review of the research, Jenkins et al. (2007) found five studies that evaluated the classification accuracy of screening measures for third and fourth grade students. In each of these studies, the screening measure used was a measure of ORF. In their review of the research, no studies of screening tools were reported beyond fourth grade. In fact, one of the most popular oral reading fluency measures, the *Dynamic Indicators of Basic Early Literacy* ((DIBELS) Good, Kaminski, Smith, Laimon, & Dill, S., 2004)), does not offer benchmark or progress monitoring passages for grades seventh to twelfth. This lack of screening materials is not due to a dearth of students who struggle with reading at these grade levels, but rather a fundamental shift in how educators currently identify and intervene with reading problems at the secondary level.

Despite the estimated "8 million youngsters between fourth and twelfth grade who struggle to read at grade level" (Biancarosa & Snow, 2004, p. 3), little research on universal screening tools for middle school students has been conducted (Allison & Johnson, 2011). Although ORF is a good predictor of reading comprehension in earlier grades, when the weight of reading success is on decoding and fluency (Schatschneider et al., 2004), as the cognitive demand of reading shifts to vocabulary and comprehension, the correlation is not as strong (Schatschneider et al., 2004; Torgesen, Nettles, Howard, & Winterbottom, 2003). According to the Final Report from Carnegie Corporation of New York's Council on Advancing Adolescent Literacy (Morsy, Kieffer, & Snow 2010), "… the major difference between reading in grades K-

5 and reading in 6-12 is the transition from *learning to read* to *reading to learn*" (p.2). This differentiation of reading tasks between elementary and middle school students, demands close evaluation of universal screening tools used to identify students in older grades. Research on the accuracy and efficiency of universal screening measures for students in the intermediate grades must be conducted.

CHAPTER 2:

Review of the Literature

This chapter begins by reviewing research on secondary students' reading difficulties. Then, the effectiveness of intervention programs with this population of students is reviewed. As an intervention framework, response to intervention (RtI) procedures are detailed. In doing so, research on universal screening measures is evaluated for accuracy, efficiency and consequential validity. This chapter is designed to provide theoretical and research foundations for the universal screening process commonly used in response to intervention (RtI) frameworks.

Reading Difficulties of Secondary Students

In early elementary school, reading instruction is focused primarily on foundational reading skills such as phonemic awareness and decoding. Universal screening tools target these reading sub-skills in the form of nonsense word fluency (NWF), oral reading fluency (ORF), and letter sound naming (LSN). Rightfully so, early identification and remediation have received a significant amount of attention and funding. For these reasons, it has been suggested that screening and early intervention should be restricted to early elementary school programs. Researches suggests, however, that students beyond early elementary school also continue to, or even begin to, struggle with foundational reading skills (Badia, 1999; Catts, Compton, Tomblin, & Bridges, 2012; Compton, Fuchs, D., Fuchs L., Elleman & Gilbert, 2008; Leach et al. 2003; Lipka, Lesaux, & Siegel, 2006; Parrila, Aunola, Leskinen, Nurmi, & Kirby, 2005). That is to say, some students who did not exhibit problems with early reading tasks will develop reading deficits as they progress through school. This group of students cannot be identified with early universal screening tools because they do not struggle with these components of reading. They will, however, need intervention to remain on track with reading skill acquisition. Researchers

(Leach et al., 2003; Lipka et al., 2006) estimate that approximately 40% of children with reading disabilities (RD) in older students have a late emerging reading disability.

At the secondary level, demands for precise vocabulary, strong background knowledge and higher level thinking skills increase. In 1999, Badian conducted a longitudinal study following a large group of students (N = 1,008) from kindergarten to seventh or eighth grades. The primary purpose of the study was to examine the stability of RD across time. Secondary analyses of the data revealed three subgroups of students: those that were identified as early poor readers (grades first-fourth), late poor readers (grades fifth-eighth) or consistently poor readers (grades first-eighth). As expected the largest percentage (6.8%) of students fell into the consistently poor readers' subgroup. This is not surprising as students who struggle with early reading skills often continue to struggle. Perhaps of greater interest however, was the differentiation between early and late poor readers. While 1.9% of the population was identified as early poor readers, 5.8% were late poor readers. This statistic supports the need to continue with RD identification and intervention efforts beyond fourth grade.

Similarly, Leach et al. (2003) categorized students with Specific Learning Disabilities (SLD) into two groups: "early identified" as at risk prior to third grade and "late identified" students identified after third grade. They found that in each group, similar student profiles occurred. Equivalent numbers of students in each group struggled with (a) reading comprehension only, (b) word reading only, and (c) both reading comprehension and word level reading. This begs the question, are these students late identified (that is we missed them with previous screening tools) or late emergent (the foundation of the reading deficit occurred beyond third grade) reading disabled? Leach and colleagues (2003) retroactively examined student data in an effort to see if somehow these students had exhibited earlier markers of potential deficits

that were missed. The late identified group's achievement scores from previous screenings were, in fact, greater than students who were identified early. In other words, the screening tools used to identify students in early elementary school were not able to isolate these students who would later struggle in reading. Previously, it was thought that the plunge in fourth grade was solely due to increased academic demands, and these deficits were limited to higher level skills associated with content area reading like comprehension and vocabulary. Leach et al. (2003) unexpectedly found, that for some students who were identified "late", reading deficits were not limited to higher level reading skills, but in fact, extended to word level reading skill deficits. The authors suggest that this provides evidence that the reading deficit is not only late identified but in fact, late emergent. This construct of late-emergent reading disabilities, supports the need to continue identification processes, through universal screening, beyond early elementary school.

Lipka et al. (2006) extended this discussion by evaluating results for 22 children who were identified with word-reading deficits in fourth grade from a sample of 1,100 children. Data were collected from kindergarten through fourth grade. Of the students identified, seven had persistent problems across all grade levels, eight students were identified after third grade (late emergent), and seven had inconsistent deficits at other grade levels. In this study, the students who were identified beyond second grade continued to display phonological awareness deficits. The authors speculated that the students in this group were able to compensate for these deficits in earlier grades, but as the reading became more challenging, their reading and spelling errors became apparent.

In a study designed to examine prevalence and heterogeneity of late emergent poor readers, Catts et al. (2012) used latent transition analysis to model changes in reading classification (good vs. poor) across kindergarten to 10th grades. They found from the sample of

493 children: 13.4 % could be classified as late emergent poor readers. Of these students, 52% struggled with comprehension only, 36% with word reading only and 12% with a combination of comprehension and word reading. This study adds to a growing body of evidence to support the existence of a subpopulation of students who cannot be identified using early universal screening tools, but will need continued support as they progress through school.

Researchers (Leach et al., 2003; Lipka et al., 2006) estimate that approximately 40% of older students with RD have a late emerging reading disability. If we recognize that all students with reading deficits will not be identified in early elementary school, then we must examine the effectiveness of intervention efforts for students in upper elementary and secondary settings.

Although progress made by struggling secondary students is not as hearty as their younger counterparts, evidence suggests students in grade four and beyond can benefit from intensive targeted interventions (Edmonds, et al., 2009; Speece et al., 2011; Wanzek, Vaughn, Scammacca, Metz, Murray, Roberts & Danielson, 2013). In a meta-analysis conducted by Edmonds et al. (2009), an overall large effect size of .89 was reported for the impact of reading interventions for decoding, fluency, vocabulary and comprehension on comprehension outcomes for older students (Grade 6 through 12) with reading difficulties or disabilities. More recently, Wanzek et al. (2013) report much lower effect sizes (.10 to .16) for reading interventions on measures of comprehension. Despite the small effect sizes reported, the authors point out,

> Nonetheless, the overall small effects noted on the standardized measures in high-quality studies illustrate that adolescence is not too late to intervene in reading and that student achievement in comprehension, word recognition, fluency, word reading fluency, and spelling can be improved in small amounts through extensive interventions. (Wanzek et al., 2013, pg. 191).

Scammacca, Vaughn, Edmonds, Wexler, Reutebuch & Torgesen (2007), also conducted a meta-analysis of reading outcomes from reading interventions in Grades fourth-twelfth. The overall effect size for the 31 studies was 0.95 ($p < .001$; 95% CI =.68, 1.22). From the 31 studies, 23 were identified with measures of reading comprehension. "With few exceptions, the pattern of results for reading comprehension mirrors the results from the overall analysis of all outcome measures. The estimate of the effect size across all 23 studies was 0.97 (95% CI.61, 1.33)" (Scammacca et al., 2007, p.8). These results indicate growth from the intervention groups that is nearly one standard deviation growth greater than the control groups. It seems evident that given appropriate instruction, students can benefit from intensive interventions even in high school. One identified moderator in this study was the grade level of the student. Effect sizes for students in the middle grades were 1.05 on all measures and 0.56 for standardized measures while results for high school students were 0.78 on all measures and only 0.13 on standardized measures. Another critical moderator considered in this study was the type of intervention given. Reading comprehension strategy interventions had a very large effect size (1.23 all measures; 0.55 standardized measures; 1.35 all measures of comprehension; 0.54 standardized measures); word study interventions had a moderate overall effect (0.60 all measures; 0.68 standardized measures; 0.40 measures of reading comprehension; multi-component interventions demonstrated a moderate overall effect (0.56 all measures; 0.41 standardized measures; 0.80 all measures of reading comprehension; 0.59 standardized measures of reading comprehension); Vocabulary interventions had the largest effect size of 1.62, however no vocabulary interventions used standardized measures; finally fluency interventions had the smallest effect size (0.26 all measures; 0.04 standardized measures; -0.07 standardized measures of reading comprehension). These results support the claim that it is never too late to provide reading interventions to

secondary students, however students in middle grades seem to respond more positively than those in high school. In addition, the type of intervention provided is also significant.

Reading deficits in secondary students are heterogeneous and multidimensional (Speece et al., 2011). Older students are more likely to struggle in more than one area (Johnson, Jenkins, & Petscher, 2010). Subgroups of students expected to struggle in later elementary identified by Johnson, et al. 2010 are: students with poor foundational reading skills, English language learners (ELLs), students requiring ongoing intervention, and/or students with late-emergent reading disabilities. Therefore, the reading developmental profiles of older elementary students require sophisticated screening tools and intervention programs to be effective. However, success of these interventions relies on systematized identification and placement of students. One popular framework currently used across the country is response to intervention (RTI).

Response to Intervention (RtI)

Response to Intervention (RtI) is a multi-tiered framework for delivering interventions to students who fail to learn at an adequate rate while engaged in high-quality, research-based instruction (National Association of State Directors of Special Education [NASDSE], 2006; Vaughn & Fuchs, 2003). Approximately ten years ago, two important pieces of legislation, the Individuals with Disabilities Educational Improvement Act of 2004 (IDEIA; 2004), and No Child Left Behind Act (NCLB; 2001) set the stage for the implementation of Rtl in many school districts across the country. RtI is designed to meet two primary purposes: first, to qualify students with specific learning disabilities (SLD); and second to provide a preventive intervention system for students at risk for academic failure.

Four key elements outlined by the National Association of State Directors of Special Education (NASDSE, 2006) for school systems to model RtI systems on include 1) design

multiple tiers of intervention and instruction where the first tier (core, or general education) meets the needs of 80% of the population; 2) provide all students with high quality targeted instruction; 3) emphasize formative assessment data to place and monitor student growth, thereby ensuring that placement continues to match appropriate instruction across time and; 4) evaluate system effectiveness across multiple tiers. A fifth element critical to the success of RtI systems is universal screening (Johnson et al., 2006).

Universal screening, in an RtI model, is the gateway to Tier II intervention for students who are experiencing difficulties in reaching reading proficiency. All students in an RtI school are regularly assessed and measured against cut scores in order to identify individual students who may be in need of intervention beyond core instruction. According to Jenkins (2003), a screening tool must satisfy three criteria: accuracy, efficiency, and consequential validity. Accurate screening measures find the correct students and only the correct students; efficient screeners are inexpensive, and easy to implement; and consequentially valid screeners have an overall positive net effect for the student (Messick, 1989).

Psychometrics of Universal Screening Measures

Accuracy. Classification accuracy (Lichtenstein & Ireton, 1984) evaluates the sensitivity and specificity of a screening measure. A high level of ***sensitivity*** assures that all students who need to be identified by the screener will, in fact, be identified. A high level of ***specificity*** assures that only the students who need to be identified will be identified. In a perfect screening tool, discrimination between those who need help and those who do not would be 100% accurate. This is, of course, an unrealistic expectation. Consider the results of a particular screening tool in two populations. One group of students is at risk, the other group of students not at risk. Rarely can

an assessment create a perfect separation between the two groups. The distribution of the test results will usually overlap, as shown in Figure 1.

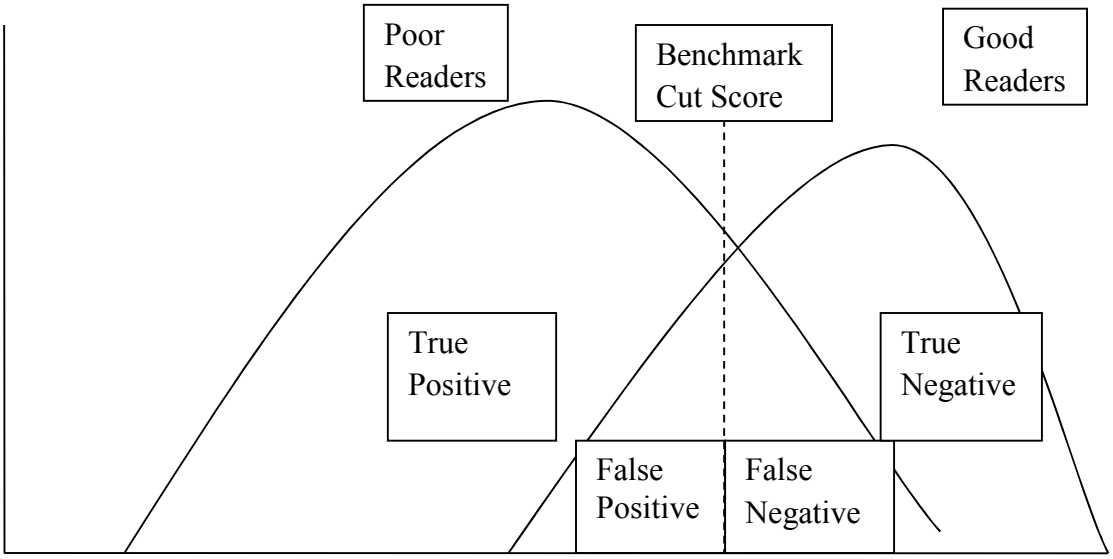

Figure 1. Illustrates potential classification accuracy results of a Universal Screening Reading Measure, where a positive result indicates a need for reading interventions.

Classification accuracy reflects the ability of the screening measure to separate these two groups as much as possible. Greater separation between the groups of students results in greater classification accuracy.

Tension between sensitivity and specificity exist. As one of these indicators increases, the natural movement is for the other to decrease. Reading researchers have agreed that the priority must be placed on maintaining a high level of sensitivity (Compton, Fuchs, D., Fuchs L., & Bryant, 2006; Compton, Olson, DeFries, & Pennington, 2002; Jenkins et al., 2007). A screening tool with 100% sensitivity identifies 100% of the students who do not pass the criterion measure. The cost of poor sensitivity is failure to identify students as “at risk” that indeed need assistance. Jenkins and colleagues (2003, 2007) suggest that 90% sensitivity should be expected from a screening tool.

Consider, however, the sobering reality. According to the National Assessment of Educational Progress, more than two thirds of the nation's fourth grade and eighth grade students are not proficient at reading (Aud et al., 2011). If we use the same proportion, in a school with 600 students, over 400 are at risk for reading failure according to NCES criteria. A screening tool with 90% sensitivity would identify 360 of these students, leaving 40 unidentified even though they need services. When Jenkins' (2003) recommendation is placed in the context of children it no longer seems unreasonable.

A high specificity level indicates that the screening measure results in few false positives. This means that the screening tool correctly identified most students "not at risk". The primary concern with low specificity is usually in the form of resource allocation. Interventions are expensive. For schools struggling to meet the needs of all students, resources are valuable. Students identified as false positive who receive interventions unnecessarily stretch an already thin budget. This can result in larger class sizes or less intervention time for students truly in need. A secondary concern with low specificity lies in intervention development. It is difficult to evaluate the effectiveness of an intervention if the students receiving the intervention did not need it to begin with. An intervention group with a large population of false positive cases may appear successful, when in fact the perception of need was merely poor performance on a single measure. It is of great importance to schools and researchers alike, that screening tools provide an accurate depiction of student need.

The diagnostic performance of a test to discriminate "at risk" cases from not at risk cases can be evaluated using Receiver Operating Characteristic (ROC) curve analysis (Metz, 1978; Zweig & Campbell, 1993). ROC curves can also be used to compare the diagnostic performance of two or more diagnostic tests (Pepe, et al., 2004). ROC curves have been used in the fields of

medicine, cognitive psychology, criminal justice, engineering and meteorology (Green & Swets, 1966; Metz, 1978; Swets, Dawes & Monahan 2000) as well as education (e.g., Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008). ROC curve is used to evaluate differences in predictive accuracy across models (Steadman et al., 2000; Tsien, Fraser, Long, & Kennedy, 1998). In a ROC curve, the true positive rate (sensitivity) is plotted as a function of the false positive rate (specificity) for different cut points. Each point on the ROC plot represents a sensitivity/specificity pair for a particular cut point. A test with perfect separation (no overlap between the two distributions) has a ROC plot that passes through the upper left corner (100% sensitivity and 100% specificity). Consequently, the closer the plot is to the upper left hand corner, the higher the overall accuracy of the assessment (see Figure 2).

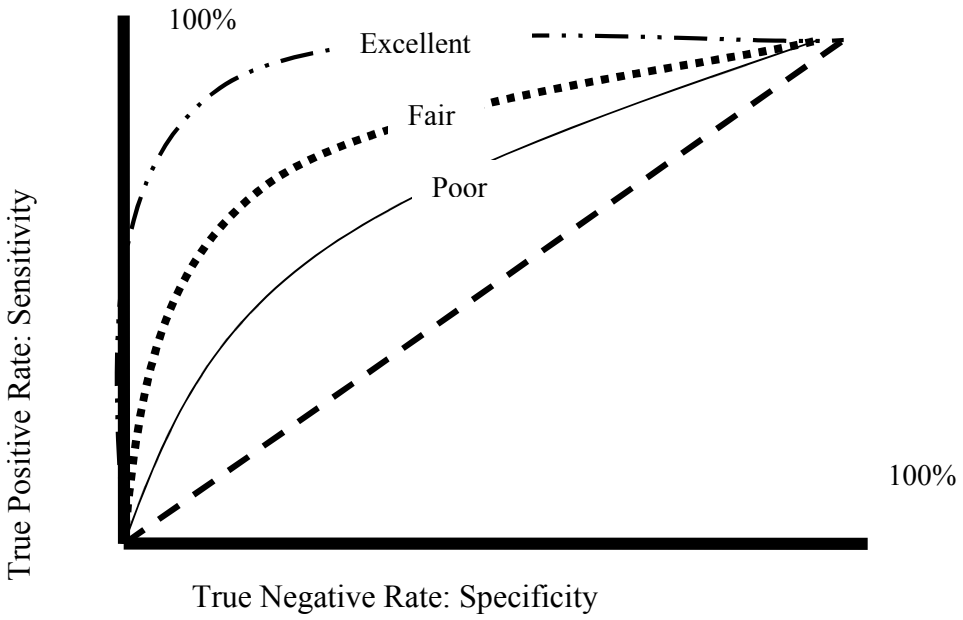

Figure 2. Examples of poor, fair and excellent Area Under the Curves (AUC) from Receiver Operating Characteristic (ROC) curve analyses.

The Area under the curve (AUC) measures accuracy in ROC curve analysis. An area of one indicates a perfect test; an area of .5 represents 50% accuracy, which is a worthless test as it

does no better than chance in discriminating between student performance (Zhou, Obuchowski, & McClish, 2002). To better understand how AUC is determined, consider the following situation. Assume that students in a classroom are accurately classified into "at risk" and "not at risk" groups. Then each pair of randomly selected students, one from each group, is tested. The student from the "at risk" group should score lower on the prediction battery. The percentage of randomly selected pairs for which this is true, represents the AUC. An AUC greater that .90 is considered excellent; .80 to.90, good; .70 to .80, fair; and below .70, poor (Compton, Fuchs, Fuchs & Bryant, 2006). Often as sensitivity increases specificity may decrease. In RtI frameworks, accuracy demands must lean toward sensitivity results. False negative errors result in a failure to intervene appropriately for students who are, in fact, "at-risk". The failure to identify students in need of intervention outweighs the need to conserve resources. As suggested by Jenkins (2003) sensitivity of .90 to .95 should be the goal for universal screening measures.

Efficiency. Efficiency refers to the perception of the users in regards to the assessment. In addition to valid and reliable, a good screener must be easy to administer and score. Simply put, it must be considered valuable by those using the assessment. Efficiency can be evaluated in several ways. First, cost/benefit must be considered in any screening instrument (Flanagan, Bierman, & Kam, 2003; Glover & Albers, 2007). Screening cost is not limited to fiscal resources, but also takes into account the human burden of administration. Assessment for the sake of assessment is not valuable. Teachers must see the impact on instruction and student outcomes in order to see benefit in lost instructional time (Bennett et al., 1999). Second, feasibility of the instrument is critical. Clear administration directions and scoring guideline should be readily available (AERA et al., 1999). Time requirements and test format weigh into the feasibility of the tool as well. Third, buy-in from multiple stakeholders is necessary. All

parties (teachers, students, parents, administrators) should see the value in the assessment. Fourth, required infrastructure to support the use of universal screening must be considered. Data collection and interpretation can be time consuming and challenging, consideration to the personnel required to maintain the model. Fifth, consideration for target populations such as students with disabilities or for whom English is a second language is essential (AERA et al., 1999). Finally, an intervention system to accommodate the needs identified by the screening process is imperative. Identification without appropriate intervention can cause more damage than good (Meier, 1975). In school-wide implementation of universal screening tools the stake-holders must value the process of screening. Research designed to ensure high accuracy standards set by Jenkins (2003) while minimizing resources (time, money, instructional loss) dedicated to the process is critical.

Consequential validity. As Valencia et al. (2010) point out, a threat to consequential validity of universal screening measures is the high rate of false negatives identified when applying commonly used benchmarks (i.e., DIBELS; Hasbrouck & Tindal, 2006) and calculations of test sensitivity (Johnson, Jenkins, Petscher, & Catts, 2009). Several studies report a high rate of under-identification when ORF is used to identify students at risk of not passing high-stake assessments. These false negative rates range from 15% to 47% (Jenkins et al. 2007; Pressley, Hilden & Shankland, 2005; Riedel, 2007; Schilling, Carlisle, Scott & Zeng, 2007). Furthermore, Valencia and colleagues (2010) also report a problem with under identification, "The reader profiles of students who were misidentified according to DIBELS and Hasbrouck and Tindal wcpm standards indicated that 30%-70% of the students demonstrated difficulty with comprehension" (p.287). That is to say, 30 to 70 percent of the students not identified (false

negative) as “at-risk”, may struggle with comprehension. Under-identification is a major concern, due to lack of intervention for students in need.

Ideally, universal screening procedures isolate academic achievement from student behavior. This separation allows for objective evaluation of individual student needs. The noisy, rambunctious boy will not receive greater attention than the quiet, solitary girl. As a result, a more equitable distribution of assistance takes place, based on academic need rather than personality traits or behavioral problems. In addition, universal screening measures do not draw attention to individual students unnecessarily, as all students take the assessment. Finally, RtI frameworks require several attempts at intervention prior to special education testing, thus avoiding special education labels. John Hattie (2009) reports in his meta-analysis an effect size of .61 on student achievement for not labeling students.

The tensions between the psychometric priorities (accuracy, efficiency and consequential validity) serve as a set of constraints for researchers in the development of universal screening procedures. Just as “Goldilocks” searches for “just right”, researchers seek the balance of quantity and variation of screening tools that will result in accurate, efficient, and valid predictions of student performance.

Although a wide variety of screening tools have been used in RtI research, few have provided adequate classification accuracy when administered in isolation (Jenkins, 2003, 2007; Johnson et al., 2010). One method used to increase sensitivity is to raise the benchmark on the screening tool. By raising the benchmark, the universal screening measure will identify more students as potentially “at-risk”. In doing so, the number of false negatives will be reduced. However, by raising the benchmark score it is likely that students who are not “at-risk” will also be identified, thus raising the number of false positives within the group. For example, Shapiro,

Solari & Petscher (2008) set sensitivity levels as close to .90 as possible (.88 to 1.0) and found ORF and 4Sight (a benchmark initiative in the Pennsylvania Department of Education) specificity ranging from .43 to .63. In this case, you can see that by moving the benchmark score up in an effort to identify nearly all of the "at-risk" students, many students who will meet benchmark on the outcome measure are also identified. Similarly, using ORF as a screening tool, Wood (2006) found sensitivity and specificity levels with the Colorado Student Assessment Program (CSAP) for third grade .86 and .64, fourth grade .95 and .58, and fifth grade .85 and .67 respectively. These classification accuracy scores illustrate the inverse relationship that exists between sensitivity and specificity. With sensitivity levels set near 90%, each of these single measure screening studies had over identification rates from 33% to57%. This level of over-identification may be unmanageable in many school settings.

Another potential strategy to improve classification accuracy is to use multiple measure screening tools (Catts, Fey, Zhang, & Tomblin, 2001; Compton et al., 2006; Davis, Lindo, & Compton, 2007; Foorman et al., 1998; Jenkins & O'Connor, 2002; O'Conner & Jenkins 1999). Within this framework, two general strategies have emerged: gated screening and screening batteries.

Gated screening method. In a gated screening method, a screening tool with a very high if not perfect sensitivity is used as the initial test. Most students who pass this test are predicted to pass the criterion measure and are eliminated from further testing. Then the remaining students are given additional testing.

Compton, et al. (2010) used this model with 712 first grade students. Students were given three 1-min fluency assessments (Word Identification Fluency, WIF N-screen; Rapid Letter Naming, RLN; and Rapid Sound Naming, RSN). Based on the results of these assessments,

students were divided in to high, average and low groups. A total of 485 children were included in an effort to increase the weight of the low group: 310 low-study-entry (LSE), 83 average-study-entry (ASE) and 92 high-study-entry. At the time of the follow-up assessment in the spring of second grade, 355 children remained in the school district and were available for testing. Compton et al. used the following measures as a first-grade prediction battery for designating risk for reading disabilities (RD): rapid digit naming (RDN), phonemic awareness, oral vocabulary, WIF, Dynamic Assessment (DA), Running Record (RR) and ORF. The measures used as a first-grade univariate screen and a second grade battery to determine RD status included: Woodcock Reading Mastery Test-word attack, word identification, passage comprehension; Test of Sight Word Reading Efficiency and Test of Phonemic Decoding Efficiency (TOWRE). Compton found that attempts to reduce the number of false positives favor assessments designed to progress monitor the response to classroom instruction such as Dynamic Assessments over assessments that measure a child's ability to read a passage with measures such as RR and ORF. In evaluating the efficacy of the 2-step procedure, the measure of phonemic decoding efficiency eliminated the greatest number of true negatives (43.4% of the sample). As a result, Compton et al. "… recommend the use of 2-step gated procedures as a means to increase the efficiency of 1-step universal screening procedures." (p 15).

Screening battery. The alternate method of increasing sensitivity in universal screening is to create composite scores from several screening measures. In this system, all students are assessed in several constructs in a screening battery during the initial screening timeframe. Although this model requires a greater time commitment during the initial screening, it provides a more complete student profile from the onset. This strategy helps identify student deficits without singling out students for further assessment.

There is evidence that a composite score results in better accuracy than the individual assessments that make up the battery. Johnson et al., (2010) found that a battery of individual screeners (ORF, Peabody Picture Vocabulary Test, and Stanford Achievement Test) used with 12,151 second and third graders were more sensitive as a single composite score than as individual scores. Sensitivity was set using ROC curve analysis to .90 for all of the comparisons. With this level of sensitivity, the specificity scores for individual tests ranged from .43 to .58 as compared to a specificity score of .68 for a probability of risk index created from the combined screening tools. Similar results for other age groups have also been found; specificity increased when a composite score was used (Compton et al., 2010; Foorman et al., 1998; O'Connor & Jenkins, 1999). Speece et al. (2010) used a slightly different technique in combining multiple screening tools. Rather than producing a composite score, Speece and colleagues identified all students scoring below the 15th percentile on any measure and flagged them as "at risk". In this manner, with sensitivity set at .90, specificity reached .63. While sensitivity scores of .90 are considered adequate, the specificity levels reported in these studies ranged between .63-.68. This over-identification rate is still potentially problematic. Reduction of over- and under-identification rates is critical to maximizing resources.

In the current study, I will use a multiple screening tool method to evaluate the accuracy of various models leading to composite scores designed to predict student performance. In this way, I will examine various combinations of universal screeners to determine the optimal set of assessments that will maintain a high sensitivity while maximizing efficiency.

Theoretical Framework

While many measures could be used as universal screening instruments for secondary students, when designing a measure or set of measures, it is important to consider the theoretical construct of reading. In her influential book, *Beginning to Read,* Adams (1990) posits a model

used to describe the processes used while reading. In this framework, she discusses the orthographic, phonological, meaning and context processors used by expert readers and under development in beginning readers (see Figure 3).We will first describe the function of each processor independently and then explore how the processors work together to construct meaning from text.

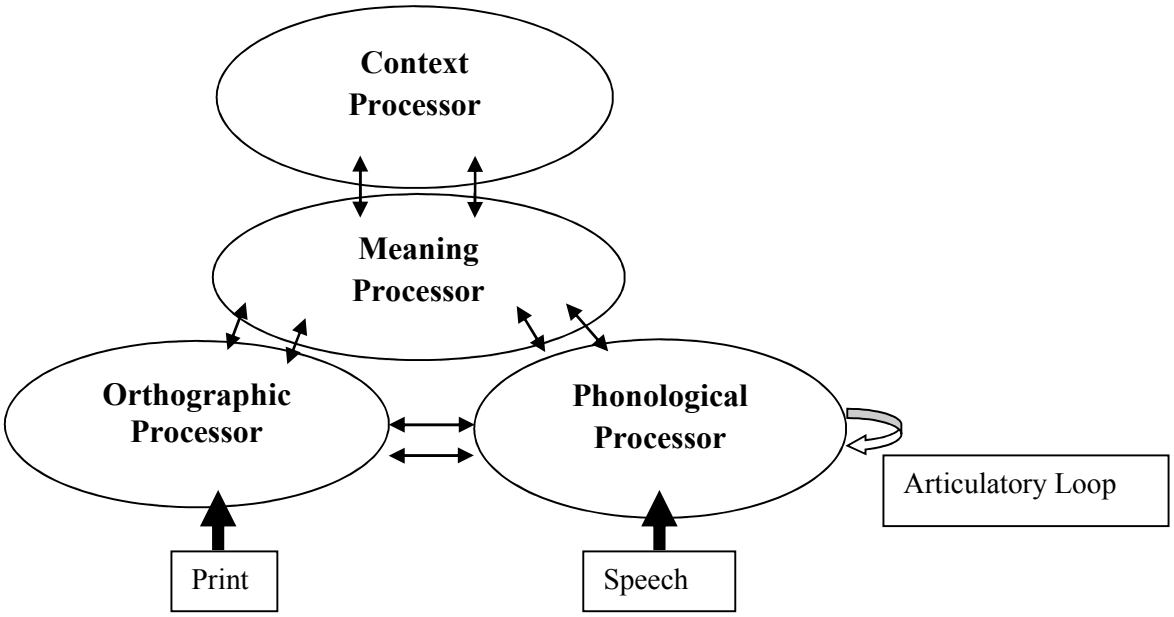

Figure 3. Adams (1990) model of the four part reading processor used to illustrate the interaction between processors as reading occurs.

Adam's Model of Reading

Parallel processing. The four-part processor framework was based on the Connectionism theoretical work of Seidenberg and McClelland (1989). As Adams (1990, p. 107) states,

> Perhaps the single most important tenant within this modeling framework is that these three types of information are not processed independently of one another. Skillful reading is the product of the coordinated and highly interactive processing of all three.

The critical nature of connectionism demands that the processors interact and develop with one another in a parallel rather than sequential manner. This feature sets the model apart from previous theories of reading. "… the parts of the reading system must grow together. They must grow to one another and from one another" (Adams, 1990, p. 6).

Orthographic processor. As seen in Figure 3, the Orthographic processor involves the intake of written information. The orthographic processor is the only processor that inputs print to the system. This seemingly simple task is actually quite complex. While this processor is responsible for letter recognition, it also is involved in the ability to see letter patterns, recognize appropriate letter order, break words into syllables and learn about likely and unlikely letter sequences. Skillful readers process individual letters in text, yet this process does not isolate letters for independent recognition but relies on associations between letters to build automaticity. The frequency of letter combinations strengthens the association and in doing so increases the stimulation to corresponding letters during the process of reading thereby creating positive excitation. In contrast letter combinations that are infrequently seen together will receive negative excitation, or inhibition, that correlates with the rarity of the combination. This negative excitation results in a reduction in fluency on the part of the reader. Take for example the nonsense word *shig,* the associations between the letter *s* and the letter *h* are strong in the English language due to the large number of words containing the diagraph *sh.* Even though the word *shig* is not part of the reader's vocabulary, skillful readers are able to decode the word fluently.

In contrast, consider the word *sgih,* in this situation the letter association are not frequent in the English language therein creating a negative excitation between the letters. Each of the nonsense words contains the same four letters, and yet fluency is greatly impacted by the order in which the letters are presented. This is partially due to frequency of the letter patterns in the words. The disruption caused by the infrequent association between letters in the word *sgih* will eventually be corrected by the direct visual information by skilled reader. In addition to the order of letters in a word, letters not included in a word that have strong association with letters in the word may also be stimulated. Take for instance the word *tae.* All three of the letters fall within the readers foveal view, that is each of the letters are seen at the same time by the reader. Because the letter combination *the* occurs so frequently in the English language the associative link between the letters *t* and *e* will excite the letter *h* even though it is not present in the word *tae*.

Letter patterns can also indicate to the reader a break in syllables. For instance, the letter combination *dn* is not part of the English language in single syllable words, but can serve as a marker for a syllable boundary such as in the word *fondness.* The inhibition between the letters *n and d* actually aides the reader in producing the correct syllable break.

To summarize, although the orthographic processor processes information regarding individual letters, it also effectively perceives whole words and syllables. "… the ability to perceive whole words and syllables as wholes evolves only through complete and repeated attention to sequences of individual letters" (Adams, 1990, pg. 130). Thus we can conclude that the development of letter and letter pattern recognition is critical to reading development. The stronger the association readers make between letters and combinations the more efficient the reader will become.

Phonological processor. Similar to the relationship between print and the Orthographic processor, the Phonological processor is the only input to the system for speech when spelling and only output from the system for reading (Adams, 1990). This processor aids the Orthographic and Meaning processors as it provides the pronunciation of words to the processors to help confirm spelling and word knowledge. Arrows between the Orthographic and Phonological processors are bidirectional, indicating that as a string of letters is being processed by the Orthographic processor; excitatory stimulation is directed to the corresponding units in the Phonological processor. If the letter string is pronounceable, the Phonological processor will send excitatory stimulation back to the Orthographic processor. This is represented in the model by the arrow running the opposite direction. Essentially, the act of decoding unknown words and encoding known words takes place during the interaction between the Phonological and Orthographic processor. The Phonological processor is also connected to the Meaning processor where the activation of a word's meaning excites the phonological units involved in the pronunciation of the word. As represented by the bidirectional arrows the pronunciation of a word also excites the meanings of the word.

It is important to note that the Phonological processor can be activated or reactivated at will. We can also speak, subvocalize or create speech images independently. The articulatory loop, indicated by the semicircle generating from the phonological processor, is a short term memory system that allows a person to hold a string of sounds or words in their head by repeating them. The articulatory loop supports comprehension by increasing the reader's verbatim memory capacity.

Meaning processor. The Meaning processor, as one might suspect, stores meaning (Adams, 1990). Perhaps of greater importance than individual word level knowledge, the

Meaning processor relies on the interconnectedness of word meanings used to build semantic webs of related words. Comparing the Meaning processor to a mental dictionary or lexicon is insufficient, as it also contains the schemas, or networks associated with that word. "… the meanings of familiar words are represented in the Meaning processor as inter-associated sets of more primitive meaning elements" (Adams, 1990, p. 143). That is to say, the Meaning processor not only contains the definition of the word, but all the associations the individual has with that particular word. Essentially, the Meaning processor represents a student's vocabulary knowledge and schemata about the world.

Similar to the Orthographic processor, the strength of the meaning processor is influenced by the number of exposures to a word. Each time an individual encounters a word; her semantic web grows and stabilizes conceptually. For instance, children encounter the word *table* relatively early in their life. They may first view the table as the thing that holds their food. Then as they start to walk they may learn that the table is hard and heavy when they bump their head. They may also learn that there are different sizes, shapes and materials used to form their concept of *table.* When a child reaches school age, they may find that a *table* is a graphical representation that holds numbers when talking about math or that there is something called a water *table* when discussing environmental issues. The word *table* has multiple definitions. The Meaning and Context processors interact with one another in order to determine which form of the word is appropriate given the particular circumstances. This interaction is represented by the bidirectional arrows included in the model.

The Meaning processor is the center of the system and is connected bi-directionally to every other processor. In such, the Meaning processor relies on the interaction between the Orthographic and Phonological processor which produces encoding and decoding. If either of

these processors is weak, it impairs the performance of the Meaning processor. The Orthographic processor generates the letter combinations and attaches these letters to the Phonological processor with the pronunciation of the word. As these processors are working together, the articulatory loop allows the reader to repeat the string of words while it connects to the Meaning processor which, as evident in the previous example of the word *table*, derives schema from the context of the situation.

Context processor. The Context processor is "in charge of constructing a coherent ongoing interpretation of the text" (Adams, 1990, p. 138). Written language does not contain several features available to the listener in oral language, such as intonation, body language, and physical surroundings. For this reason, sentence contextual clues are far less effective for acquiring vocabulary than one might think. The Context processor does, however, rely on sentence structure to help synthesize meaning from a written passage.

The Context processor is responsible for selecting between multiple meanings of words. It does not however prevent excitation to inappropriate uses of the word. Consider the sentence, *Jo found the table included in the report confusing.* While the reader should quickly settle on the definition of the word *table* as a graphic representation, they will also briefly, identify the use of the word with the piece of furniture. The speed at which this occurs is due to the efficiency of the Context processor. This phenomenon also demonstrates the strength of the Orthographic processor. Although the Context processor is able to identify the correct word usage, it is unable to prevent the Orthographic processor to excite each of the possible meanings. This leads us to the conclusion that although the Context processor can assist the Orthographic processor, it cannot replace it. In addition, this also may reveal potential confounding factors for English Language Learners (ELL). For beginning or struggling ELL students, the Context processor will

not be able to assist the Orthographic processor in deriving meaning from text as the associations for multiple word meanings is impaired due to language acquisition.

A sentence that is highly supported by context is strongly predictive of the word to follow. This will allow the Context processor to send a strong and focused excitation to the Meaning processor. This predictability of words allows the fluent resolution of word choices as the string of letters is being processed. In essence, the Context, Orthographic and Phonological processor are all feeding information to the Meaning processor in unison. The greater the excitations, created by exposure and predictable patterns, the faster the Meaning processor are able to generate the appropriate word.

Reading fluency. A critical aspect of the Adams (1990) model is the role automaticity plays in overall reading comprehension. As we have discussed, each of the processors increases the speed in which they pass information to other processors through excitations developed through multiple exposures with sounds, letter patterns and conceptual understandings. An essential component of the overall functioning of the model is fluency. As Adam's (1990, p. 160) points out,

> …the utility of the associative linkages, both within and between processors, depends on the speed and completeness of the input they receive. When the words of a text are processed to slowly or scantily, readers forfeit any automatic facilitation and guidance that the associative connections would have otherwise provided. Commensurately, they also forfeit the opportunity to recognize, learn about, and understand what they have read.

As discussed earlier, positive excitation between the four processors over time bolsters connections for particular constructs and therein increases the speed at which the brain makes

these connections. This is represented in Figures 3 and 4, with the bidirectional arrow that extends across the length of the four part processing system. This representation suggests the importance of fluency throughout the system yet maintains this as a separate construct. The stronger the neural connections between processors, the faster students are able to respond to prompts connected to that processor. Measurements of fluency such as NWF, LNF and ORF, leverage this phenomenon with very short assessments (one to three minutes in length) that quantify the speed at which a student can provide a response. The assumption is, the faster the response, the stronger the neural connection between the processors involved in the task.

Strong reliability and validity of CBM oral reading fluency passages has been established over the past twenty years (e.g. Good, Simmons, & Kame'enui, 2001; Marston, 1989; Shinn, 1989; Wayman et al., 2007). Marston (1989), in a review of 11 studies, reports a mean reliability of .91 (SD = .04). In the *Becoming a Nation of Readers: The Report of the Commission on Reading,* Anderson (1985) recognized: "A more valid assessment of basic reading proficiency than that provided by standardized tests could be obtained by ascertaining whether students can and do the following: Read aloud unfamiliar but grade-appropriate material with acceptable fluency…" (p. 99).

Stage & Jacobsen (2001) conducted a study with 173 fourth-grade students. Classification accuracy for oral reading fluency (ORF) was evaluated for two cut scores (< 100, < 50 wcpm). The criterion measure in this study was "meets expectations" on the Washington Assessment of Student Learning (WASL). Using < 100 wcpm as the cut score, sensitivity and specificity were 66% and 76% respectively. By lowering the cut score to < 50 wcpm, specificity was increased to 96%. However, this increase came at a cost to sensitivity, which dropped to 31%. McGlinchey & Hixson (2004) evaluated a larger sample (n = 1362) of fourth grade

students and found slightly higher classification accuracy levels of 75% sensitivity and 74% specificity when using ORF to predict the Michigan Educational Assessment Program (MEAP).

One potential explanation for the difference in classification accuracy between the two studies could be time of administration. Data gathered for the Stage & Jacobsen (2001) study was gathered in the fall, while McGlinchey & Hixson (2004) gave the ORF screen in the winter. Theoretically, the closer administrations of the screen and criterion measure are to one another, the easier it is to predict the outcome; therefore the classification accuracy should be higher. Regardless, neither of these studies approach the 90% sensitivity level recommended by Jenkins et al. (2007).

Conceptual Framework

Due to the interdependent nature of the four-part processors in the Adams (1990) model, weakness in one processor impacts overall reading performance. When thinking about screening assessments and detecting students at-risk for reading problems, consideration of each of the processors may prove valuable. Some students may be able to compensate for impaired early reading skills through overreliance on alternate processors. For example, a student with a phonological processing deficit who possesses a large vocabulary corpus and mature background knowledge schema may depend on the Context and Meaning processor to aide in word recognition. While the Context processor cannot replace the efforts of the Phonological and Orthographic processors it may be able to, at least in early reading tasks, provide support. Similarly, ELL students with limited contextual schema may develop early decoding skills rapidly. The Orthographic and Phonological processors may compensate for the Meaning processor in measures of early reading skills. Or perhaps more accurately, the measures of early reading skills (nonsense word fluency (NWF), letter naming fluency (LNF), phoneme

segmentation fluency (PSF)) may include tasks that do not require the Meaning and Context processor to contribute to the task.

If the strength or weakness of each processor is isolated through screening measures designed to detect individual deficits, students who compensate for weak processors may be more accurately and efficiently identified as "at-risk". Although it would be difficult to tease apart the processors completely, there are measures that rely more heavily on individual processors or the bidirectional relations between two processors than others. For instance: nonsense word fluency, word identification fluency, and phoneme segmentation fluency are all common screening tools used with students in early elementary school. These foundational reading skill assessments are heavily weighted with information regarding the Orthographic and Phonological processors and may only indirectly tap into the Meaning and Context processors. Students with vocabulary or background knowledge deficits are less likely to be identified using these tools.

In order to develop a potential set of screening measures appropriate for older readecrs, I developed the following proposed assessment model where I match screening measures to portions of the four-part processor (see Figure 4). In this model, the ovals represent the form of language input (oral vs. written); while the rectangles represent potential screening measure constructs.

Screening Constructs

The screening constructs used to measure "at-risk" status of student in middle grades in this model are designed to mirror Adam's model. In the theoretical framework for this study I looked at the Orthographic, Phonological, Meaning and Context processors and fluency respectively. I mirror this order of discussion in the conceptual framework by discussing the

constructs of reading that are most heavily relied on for each of these processors: encoding and decoding, vocabulary, comprehension and fluency. While I recognize that all processors are working together in the act of reading, it is possible that weaknesses can be detected in portions of the system with specific assessments.

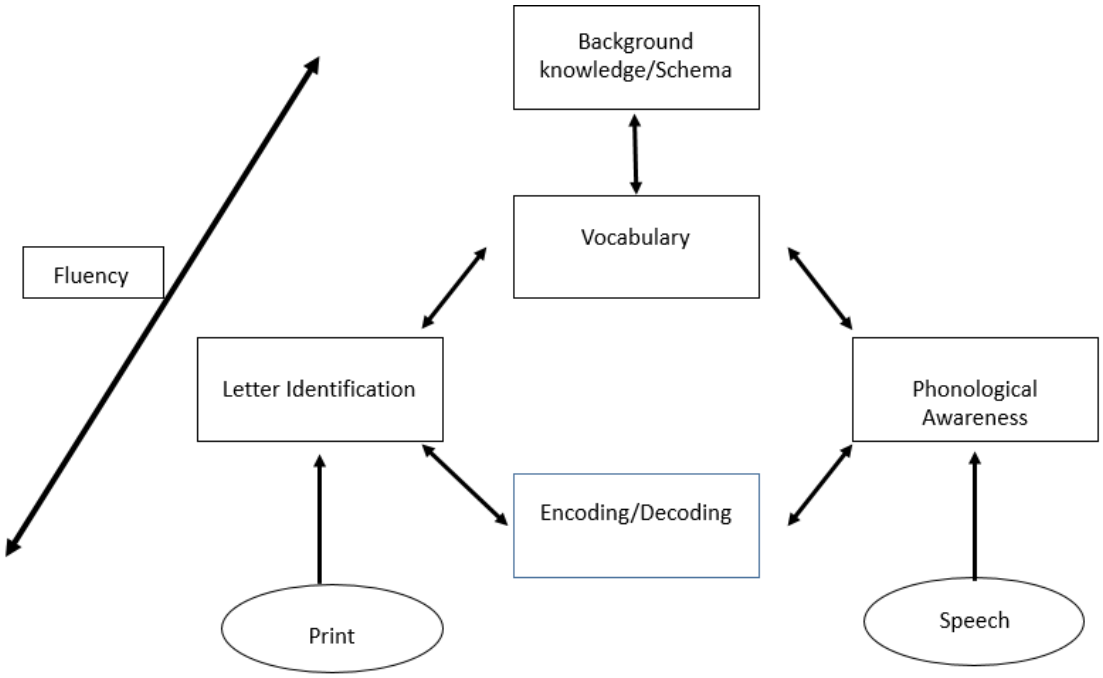

Figure 4. Reading assessment model

Encoding and decoding. The interaction between the Orthographic and Phonological processors is primarily responsible for the encoding and decoding. That is to say, a problem in either one of these processors would result in a student's difficulty to decode (read) or encode (spell). While measures of decoding are commonly used with elementary school students, it is very time consuming to administer these assessments. Because the students in this study are in sixth grade, the district from which the extant data was obtained, did not feel that one-on-one administered decoding assessments was viable. Instead they used a possibly more developmentally appropriate spelling assessment to measure encoding. Some students compensate for a deficit between the Orthographic and Phonological processor interaction and

can reach benchmark in measures of silent reading. Spelling screeners, however, can often detect the "at-risk" students (Foorman & Ciancio, 2005; Speece et al., 2010) missed by silent reading screening tools.

Little research has been conducted with spelling as a predictor variable for overall reading comprehension. However, when we examine Adam's (1990) model, we can see the potential. For older students, spelling can often be used as a proxy for decoding skills. The act of spelling often begins in the phonological processor where the student hears a word that he/she is then asked to spell. This receptive oral skill requires that the student hear the word, identify the phonemes in the word, and finally map the appropriate graphemes to the phonemes in the word. While engaging in this process, the brain is also accessing information from the meaning processor, such as morphological information about the word parts. In addition, the context processor is assisting the process if the word was given in the context of a sentence. The context processor may be sifting through alternate spellings based on this information. Although spelling often begins in the phonological processor or context processor, you can see how all the processors aide in the construction of a word. Spelling can also be generated from the meaning processor as a student selects a word that they want to use when writing independently. In this case, the student begins with the context processor and works their way down to the orthographic processor. This represents the expressive form of language, writing. It is also interesting to note that spelling is initiated as an orally receptive process in the input of information and ends as a written expressive process on output.

Foorman & Ciancio (2005) suggested that spelling may serve as a potential reading screener in their study of 3rd grade students. Using multiple regression, they found that spelling

accounted for 53% of the unique variance and text fluency accounted for 27% of the unique variance on an end of year outcome measure, the WJ-III Broad Reading cluster.

Speece et al. (2010) used a spelling fluency measure as one of the potential screening tools in their study of 4th grade students. Spelling Fluency (Fuchs, Fuchs, Hamlett & Allinder, 1991) is a group-administered test where for two minutes, every 10 seconds, students are presented a randomly selected word from the Harris- Jacobson grade-level list. Spelling Fluency was one of twenty-seven predictor variables considered for inclusion in this classification accuracy study. Of interest, CBM spelling was in three of the top six combinations of screening tools. In addition, the combination with the highest R^2 value (R^2=0.49) was CBM Maze, CBM Spelling and Teacher Reading Rating. From this analysis it seems reasonable to conclude that CBM Spelling may, in the future, be a strong candidate for membership in a reading screening battery.

Vocabulary. Moving up the assessment model, (Figure 4) standard measures of vocabulary knowledge seem to tap into the primary efforts of the Meaning processor. Although the Meaning processor is responsible for an expansive web of interconnected concepts, simple measures of vocabulary knowledge could gauge the overall maturity of this processor sufficiently in a screening measure.

Vocabulary assessments often require students to read a portion of text and select the appropriate word for that context. This can be done in a variety of ways. Students can be asked to select between words, as in a multiple choice or word bank model. A more rigorous method of measuring vocabulary knowledge is to require production of vocabulary words where students are not given choices but must generate an appropriate word choice. This is often a model where students must fill in the blank within a sentence. While the first method of selecting of words is

not as difficult as generating words, it may have greater potential as a screening tool. Several limitations of word selection measures are actually strengths in regard to screening tools in that they are easy to score and do not require judgment on the part of the administrator, thereby increasing inter-rater reliability.

Research on the importance of vocabulary knowledge for school success, and reading comprehension in particular, is plentiful (Anderson & Nagy, 1991; Baker, Simmons & Kame'enui, 1998; Becker, 1977; Cunningham & Stanovich, 1998). According to Coyne, Kame'enui, and Carnine (2007), "The learning characteristics that have the strongest causal connection to academic failure are rooted in the area of language" (p. 38). A strong link between vocabulary knowledge and reading competence has been long established through correlational and factor-analytic studies (Anderson & Freebody, 1981; Beck & McKeown & McCaslin, 1983; Davis, 1944, 1968; Singer, 1965; Thurstone, 1946). Sternberg (1987) has argued, in fact, "one's level of vocabulary is highly predictive, if not determinative, of one's level of reading comprehension" (p. 90). Experimental evidence also suggests that vocabulary knowledge influences comprehension (McKeown et al., 1983).

Research is beginning to suggest that vocabulary may add value in predicting student performance when combined with other screening measures. Initial studies have primarily been conducted with younger students (first through fifth grade); I extrapolate from this research that vocabulary may be a useful construct to consider for older students as well. As students develop, vocabulary plays a larger role in reading and therefore should be included in universal screening frameworks for older students.

Schatschneider, et al. (2004) evaluated approximately 200 children in grades three, seven and ten, on tests designed to measure five broad areas of reading: verbal knowledge and

reasoning, text reading fluency, phonemic decoding efficiency, non-verbal reasoning, and working memory. They sought to identify skill deficits in students who scored below benchmark on the Florida Comprehensive Assessment Test (FCAT). Of interest, in third grade students who scored in level one or two, demonstrated a weakness in ORF. In addition, third grade students who were in level one had a deficit in phonemic decoding. Verbal reasoning for third grade, however was at the 42nd percentile. This profile is significantly different for students in seventh and tenth grade. While students in seventh and tenth grade at level one continue to struggle with ORF (seventh and eighth percentile respectively) and phonemic decoding (27th and 18th percentile respectively), verbal reason also begins to drop as students get older. Where third graders were at the 42nd percentile for verbal reasoning, by seventh grade they were down to the 34th percentile and in tenth grade, the 30th percentile. This suggests that for the students who struggle the most on the high stakes state assessment demonstrate a stable deficit across time for ORF; as they grow older, verbal reasoning also begins to drop. With this drop in verbal reasoning, may come an additional construct to aid in prediction of reading failure.

Riedel (2007) found that for first grade students whose oral reading fluency was adequate, but reading comprehension was not, vocabulary assisted in more accurate identification of the students who were at-risk. Similarly, Johnson et al. (2010) included the Peabody Picture Vocabulary Test (PPVT) in a screening battery with ORF and the SAT-10 with third grade students. In this study, when considered as a single predictor, classification accuracy for PPVT (sensitivity .90, specificity .47) was as high as ORF (sensitivity .90, specificity .43). Grouped together as a screening battery PPVT, ORF and SAT-10 reached even greater classification accuracy (sensitivity .90, specificity .68). This supports the idea that vocabulary may help identify students who read fluently but fail to comprehend adequately.

Van Hook (2008) included a different measure of vocabulary knowledge in his study with first, third and fifth grade students. This multiple choice test asks students to look at a picture and then identify the word that best matches that picture. This is a three minute group administered assessment. The student score is the number of items answered correctly. In this study, Van Hook found similar classification accuracy for the Picture Word Fluency assessment as passage and sentence Maze with the GRADE (1st grade) and Mississippi Curriculum Test (third and fifth grade), with sensitivity ranging from .80 to .83 and specificity from .71 to .88. Unfortunately, the base rate in this study was very small ranging from three to eleven percent. However, this study is one of the first attempts to conduct a group administered vocabulary test as a screening tool in RTI. Not only is this tool quick to administer (about five min.), but it is also easy to grade (multiple choice) and interpret (raw score). This type of tool may hold promise in the future.

Adlof, Catts and Lee (2010) found that for eighth grade students, the model that produced the highest AUC value (0.868) was an eight-predictor model containing phoneme deletion, grammatical completion, nonverbal intelligence, and sentence imitation, mothers' level of education, narrative expression, narrative comprehension and oral vocabulary. The inclusion of oral vocabulary in this study indicates that further research should be conducted with this construct.

Finally, Nese et al. (2011) conducted a study, with 1,800 fourth and fifth grade students, which evaluated the diagnostic efficiency of easyCBM measures of ORF, vocabulary, and multiple choice comprehension. They also examined student demographics and prior year performance in predicting passing status on Oregon's high-stakes assessment. While prior year assessment explained 70% of the variance on the outcome measure, easyCBM ORF, vocabulary

and multiple choice reading comprehension were all significant with easyCBM Vocabulary having the largest effects of the three ($\beta = 0.24$ and $\beta = 0.22$ for fourth and fifth grade respectively). Again we see the potential of vocabulary measures in predicting future reading performance.

Measures of vocabulary knowledge, even the less rigorous word selection measures, require competence in each of the four processors. The Phonological and Orthographic processors must interact to decode each of the words. At the same time the Meaning and Context processor interact to make a selection from options of vocabulary words. One might argue that measures of vocabulary are often actually measures of overall reading comprehension (albeit limited ones), as they require the engagement of all four of the processors.

Comprehension. More standard measures of comprehension attempt to capture the culmination of all the processors working together in an effort to express meaning. All the processors are actively engaged. Automaticity of each of the processors plays a substantial role in the ability of the student to construct meaning from text. Assessments of comprehension vary in length and difficulty. This variance is important to recognize when evaluating assessment results. Many comprehension measures are timed measures that require students to read a sentence and make a decision based on the information in the sentence.

For instance, Maze passages delete every seventh word in the passage and provide three options for a student to choose from. While these measures are quick and simple to administer and provide easy to interpret data, they are time consuming to score. The *Test of Sentence Reading Efficiency* ((TOSRE (Vocab + Comp)) Torgesen, Rashotte, & Pearson, 2010) on the other hand, requires students to read a sentence and then respond yes or no. This assessment has similar benefits as the Maze passages, but is also easy to score. While these measures are

attractive as screening choices, the timed nature of the assessment may be misleading for students who are simply slow processors. These students may be identified as false positives due to the fluency consideration of the assessment. Measures of comprehension that require more time, often ask the student to read a passage and answer multiple choice questions. While these assessments more closely mirror tasks associated with common outcome measures, they require greater assessment time. Consideration of testing time requirements is a critical feature of effective screening models.

Standardized group administered comprehension tests such as the Stanford Achievement Test -10 (Johnson et al., 2010), Gates-MacGinitie Reading Test (Speece et al., 2010), and 4Sight Benchmark Assessment (Shapiro et al., 2008) have recently been used in screening research for students in fourth through tenth grade. Although, these assessments often do not meet the requirement of being quick to administer, as Jenkins et al. (2007) recommends, they are group administered and fairly easy to score. These longer assessments also more closely resemble outcome measures such as high stakes state assessments. The Broad Screen (Torgesen, 2003) is in fact a computer adaptive multiple choice test specifically designed to mirror the FCAT (Florida Comprehensive Assessment Test). This assessment takes between 10-30 minutes and is used as the first gate in a gated-screening procedure.

While measures of comprehension potentially offer a closer approximation of the outcome measures they strive to predict, they also are generally require more student time in assessment than is desirable of universal screening measures. Group administered comprehension measures increase student time engaged in assessment and thereby reducing time available for instruction. This tension between assessment and instruction is a large consideration when selecting screening measures.

Reading Fluency. As discussed previously, the interaction between the Orthographic and Phonological processors are primarily responsible for decoding. When the reader has strong excitations associated with letters and letter patterns in the Orthographic processor and with the speech sounds and pronunciation with the word in the Phonological processor it is likely that this automaticity will be reflected in measures of fluency.

Reading fluency assessments are desirable as universal screening tools as they are quick to administer (usually between one and three minutes), easy to score (quantitative in nature) and simple to interpret. Many fluency measures are, however, not group administered. The one-on-one nature of these measures taxes resources and is time consuming for the administrator of the assessment. While group-administered fluency measures such as the *Test of Silent Word Reading Fluency* (Speece, 2012), Sentence Maze (Allison & Johnson, 2011; Van Hook, 2008), and Passage Maze (Van Hook, 2008) satisfy the previously mentioned requirements, they remain a more distal assessment of comprehension than the previously mentioned assessments of vocabulary and comprehension.

Traditional screening tools, such as Oral Reading Fluency (ORF) are strongly correlated with overall reading ability in first to fourth grade students (Deno, Mirkin, & Chiang, 1982; Shinn, 1989). As students move to the intermediate grades, this correlation drops off significantly (Jenkins & Jewell, 1993; Schatschneider et al., 2004; Silberglitt, Burns, Madyun & Lail, 2006). Several factors may contribute to this shift. Instruction may remediate basic skills, students may learn how to take the test, or other elements of reading may begin to cause new significant problems. As Badian (1998) said, “as the nature of reading changes, so change the predictors” (p. 478). This is problematic, as the number of screening measures designed for intermediate students is extremely limited.

In a review of screening literature, Jenkins et al. (2007) identified thirteen studies. None of these studies included students beyond fourth grade. Of the third and fourth grade studies, ORF was the only screening measure used. Late emergent RD students often do not struggle with decoding or reading fluency and therefore may not be identified using ORF. Badian (1999) and Catts and Hogan (2002) stated that students with late-emerging RD in second to fourth grade possessed average fluency and word identification skills, but struggled later with reading comprehension. On the other hand, as mentioned earlier, Leach et al. (2003) identified multiple reading profiles for late emergent RD with: 36% poor decoders; 32% specific comprehension deficit and 32% exhibiting difficulty with both decoding and comprehension. Lipka at al. (2006) found yet another pattern of student deficit for late emergent RD with 68% experiencing a decoding deficit and 32% had a specific comprehension deficit. Although the breakdown of student deficit differs, it appears that there are a percentage of students who do not struggle with early reading skills that do struggle with reading competence in later elementary school. Therefore choice of universal screening measures should reflect this finding utilizing multiple components of reading to maximize classification accuracy

Torgesen, Nettles, Howard, & Winterbottom (2005) conducted a study with fourth, sixth, eighth and tenth grade students examining the relationship between the screening measures of oral reading fluency (ORF), MAZE, Test of Sentence Reading Efficiency (TOSRE (Vocab + Comp)), Test of Silent contextual Reading Fluency (TOSCRF) using the Florida Comprehensive Assessment Test (FCAT) as the outcome measure. At fourth grade they found no important differences in correlations of the screening measures with the FCAT ranging from .48 to .56. At sixth grade the MAZE test seems to more accurately predict performance on the FCAT with correlation rates of .67 than does ORF (.59) or TOSRE (Vocab + Comp) (.58). The eighth grade

results had all measure with similar correlation rates for ORF, MAZE and TOSRE (Vocab + Comp) (Vocab + Comp) .62, .63 and .58 respectively. Finally, at tenth grade ORF and TOSRE (Vocab + Comp) correlate similarly at .55 and .56 while MAZE seems less reliable at .32. The mixed results of this report seem puzzling at first. Although correlations between ORF, MAZE and TOSRE (Vocab + Comp) with the FCAT seem to fluctuate across grade levels, these correlations remain moderate between .48 and .67. This study did not look at sensitivity/specificity, positive predictive power/negative predictive power, or ROC analysis therefore the findings are limited. This study does however suggest that further evaluation of screening tools is needed for students across the developmental trajectory.

Given the wide use of ORF measures, comparisons with other group-administered measures of fluency are necessary. The greatest strength of fluency measures is, of course, the limited amount of time it takes to administer. This minimizes the time spent away from instruction. While this is a desirable characteristic, effectiveness of the tools must meet classification accuracy requirements. As students develop, the burden of the processors shifts from the lower portion of the model (Orthographic and Phonological) to the higher portion of the model (Meaning and Context). This shift may result in lower classification accuracy rates for distal measures of fluency.

Screening Measures Summary. As Compton et al. (2008) suggest, "Given the rapid development of reading skill during elementary school, it is important to select latent class indicators (i.e., measures of reading skill) that tap the critical skills aligned with the developmental phase of the readers" (p. 330). Researchers suggest that for late elementary and middle school students, vocabulary or comprehension (Compton et al., 2006; Davis et al., 2007; Jenkins et al., 2007; Johnson et al., 2009; and Riedel, 2007) may be appropriate measures. In

addition, Spelling (Foorman & Ciancio, 2005; Vaughn et al.,2010), teacher rating (Elliott, Huai, & Roach, 2007; Jenkins, 2003; Davis, et al., 2007; Ritchie & Speece, 2004; Speece et al., 2010; Vaughn et al.,2010; Fuchs, L., Fuchs, D., & Compton, 2010); and prior state assessments (Allison & Johnson, 2011) may all hold promise as screening measures for older students. Fortunately, these higher level skills lend themselves more readily to group administered screening which may ease the burden placed on teachers and school systems.

Purpose of the Study

The purpose of this study is to examine the psychometric characteristics and classification accuracy of multiple universal screening measures for secondary students while maintaining a reasonable demand on resources.

This study extends understanding of universal screening from the existing research in primary grades to middle school students. The psychometrics and relative classification accuracy of multiple screening measures from two categories: (a) demographic data such as gender, special education status (SPED), free and reduced-price lunch status (FRL), English language learners (ELL) and prior year high stakes assessment results and (b) common reading screening measures such as oral and silent reading fluency, vocabulary, and comprehension will be examined.

Research Questions

This study evaluates the following research questions:

1. What are the direct, one-to-one relationships among 2010 demographic, prior MSP, and reading screening measures with 2011 MSP Reading passing status (i.e., when children are in sixth grade)? In other words, do we see significant relationships among all measures? What is the strength of those relationships?

2. What are the sensitivities and specificities of each *individual* reading screening measure with 2011 MSP Reading passing status, *without adjustment* for the prior year MSP score or demographic characteristics?
3. What are the unique contributions of student demographic characteristics and prior year MSP Reading score to predicting 2011 MSP Reading passing status (henceforth "baseline covariates")?
4. What is the contribution of each *individual* reading measure to predicting 2011 MSP Reading passing status, *after adjustment* for (above and beyond) baseline covariates?
5. What are the contributions of the combined set of reading measures to predicting 2011 MSP Reading passing status, after adjustment for (above and beyond) baseline covariates?
6. What are the individual and combined contributions of all reading measures to 2011 MSP Reading passing status, *without prior year 2010 MSP Reading scale score*? In other words, what does the model look like for children who might not have been assessed on the prior year test, such as students who move frequently?

CHAPTER 3:

Research Methods

Setting and Participants

The data for this study are from anonymized extant data. Students attended a single middle school in a small suburban school district in the Pacific Northwest. The district comprised 5,000 students, distributed over six public elementary schools, one middle school, and one high school. In the school where this research was conducted, screening data are collected as part of the regular educational process twice a year in the fall and spring. This data and the process by which it was collected were made available to the researcher with no identifiable markers for the purpose of dissertation research.

The pre-existing database used in this study includes all sixth grade students ($N = 344$) in the sole district middle school. Of the 344 students, 195 (57%) were male, 48 (14%) had active IEPs for special education services (disability status was not reported for one student), and only 8 (2%) received English Language Learner services. Low socio-economic status was determined by eligibility for free and reduced-price lunch: in the sample, 193 students received FRL (56%) (FRL information was not reported for two students) District records on student ethnicity and race were available for all but one student. The sample distribution included 201 Caucasian students (62%), 26 Black (8%), 11 Asian (3%), 24 Hispanic (7%), 8 American Indian/Alaskan Native (2%), 9 Pacific Islander (3%), and 46 Multi-Racial (14%).

Measurement

Procedures. First, of note, all data used in this analysis are extant data collected by the school district as part of their universal screening process. Details regarding procedures were

reported to the researcher to help establish reliability and validity of the data collected. None of the assessments were conducted, scored, or recorded by the researcher.

Students were tested in their homeroom class by their teacher. This 45-minute long nonacademic class has approximately 20 students per instructor. A two week testing window is established school wide at the beginning of the school year. ORF is the most time consuming of the measures as it is an individually administered assessment. This assessment takes teachers between four and five days to complete. The remaining assessments are given with the following time frames: Day one, Test of Sentence Reading Efficiency; Day two, Vocabulary. Two days were used as make-up days at the end of the two-week period.

Teachers received a Teacher Guide at the beginning of the assessment period that includes the schedule, answer keys, test administration directions, and grading procedures for each of the assessments. As the teachers gave the assessments, they graded and inputted the data into a spread sheet. Three instructional coaches were available during this period to answer questions or provide assistance in the testing process.

The Gates-MacGinitie (Vocab + Comp) was administered in the fall and spring by the language arts teachers during the students' regular language arts period. This is a timed assessment and the teachers involved received training and administration assistance. In addition, students were given the reading portion of the Measurement of Student Progress (MSP) in May proctored by their language arts teachers.

Training. As reported by the school district, all teachers in the middle school are trained to administer the screening battery of test measures used in the current study. All teachers in the school participated in the screening process. Of the 54 teachers at this school, 51 participated in all training opportunities. Three teachers were new to the school in the second year of

implementation and there were no new teachers to this school in the year data was collected for this study.

Child demographic characteristics

In the future, comprehensive RtI models may consider multiple risk factors beyond academic performance. These factors are not present in the conceptual model presented previously, but warrant consideration. Among the most important factors affecting child development is family background. Poverty, mother's level of education, and low parental occupational status have been linked consistently with lower cognitive and social performance of children (Bradley & Corwyn, 2002; Dearing, Berry, & Zaslow, 2006; Duncan & Brooks-Gunn, 2000; Garbarino & Ganzel, 2000; McLoyd, 1998; Pike, Iervolino, Eley, Price, & Plomin, 2006). As Jenkins et al. (2007) suggest, other risk factors are potentially valuable. Often this information is readily available and requires no testing resources to accumulate. Future research in reading screening may consider investigating into predictability of these factors as part of a screening model.

Reading measures

The measures used in this study reflect the assessment model (Figure 4) in an effort to tap into information regarding potential problems in the four processors previously discussed. In addition, assessments were selected based on the limited research available on universal screening for older students (McGlinchey & Hixson, 2004; Shapiro et al., 2008; Speece et al., 2010; Stage & Jacobsen, 2001) as well as a review of the National Reading Panel (2000) components of reading and the Conceptual Framework described in Chapter 2. See Table 1 for a summary of the measures used in this study.

Table 1. Data Collection Summary

Construct	Measure	Collection Occasion	Analytic Function
Demographic Characteristics	School Records	Fall 2010	Predictor
Reading Comprehension Achievement	Grade 6 Measurements of Student Progress (MSP)	Spring 2011	***Outcome***
	Grade 5 Measurements of Student Progress (MSP)	Spring 2010	Predictor
Vocabulary	CBM Vocab	Fall 2010	Predictor
Reading Fluency	AIMSweb Oral Reading Fluency (ORF)	Fall 2010	Predictor
Vocab + Comp	Gates-MacGinitie (Vocab + Comp)	Fall 2010	Predictor
	Test of Sentence Reading Efficiency (TOSRE (Vocab + Comp))	Fall 2010	Predictor

Prior year state assessment (Reading achievement). Because state assessments generally begin in third or fourth grade, one logical source of information regarding future student performance may lie in previous state assessment scores. The value in using extant data, of course, is that it eliminates additional assessment time requirements. By sixth grade we often have substantial evidence that a student is in the highest and lowest quadrant of the group. That is to say, we can predict from prior year assessments the students who are at high risk from the students who are at low risk. The difficulty lies with those who fall in the some risk category. These students are often referred to as the "bubble kids" as they fall within range that can place them below or above benchmark on the outcome measure. For these students, it may be difficult to predict future benchmark performance based on prior year assessment.

Allison and Johnson (2011) used seventh grade scores from Georgia's Criterion Referenced Competency Test (CRCT) to predict eighth grade performance on the CRCT. This study found that ORF and Maze did not significantly improve identification of "at-risk" students over the previous year's CRCT scores alone. CRCT-7 had an AUC value of 0.96 with sensitivity and specificity at 71% and 96% respectively. These results must be viewed with caution, however, due to the very low base rate (3.4%) of the population.

Inclusion extant data as part of a screening battery, or perhaps as the first gate of a gated screening program, holds promise. In order for this to be a viable option as a first gate, the sensitivity level would need to increase well above 0.90. Perhaps with the risk criterion set at a level four (exceeds state standards) instead of level three (meets state standard); the sensitivity of the tool would increase to a desirable level. In the state of Florida, 35% of students scored at level four or level five in 2010. Hypothetically, if the classification accuracy were acceptable using this cut score, up to 35% of the students could be eliminated from further testing state-wide. As the first gate in a screening system, these data already exist and do not require any additional testing.

The main benefit of this strategy (extant data) may, in fact, be the main detriment as well. Because the test is only collected once a year, students who were not present will not have a score and cannot obtain one until the spring of the following year. If the universal screening model relied on this measure as part of a composite score, this could present logistical concerns. If, however, the state test served as the first gate of a screening model, new students without state scores could be directly placed into the second gate of the screening system. In either case, schools with high transient populations may find state testing as a screening measure difficult to manage.

Reading Comprehension Achievement. In this study, the focal measure of reading comprehension achievement used in this study is the Washington State *Measurement of Student Progress* (MSP), which is a battery of state-level assessments designed to measure student progress toward mastery of the Essential Academic Learning Requirements (EALRs) for the state of Washington. It is given each April-May. In the first year of its implementation, the MSP tested approximately 25% of the students in the state of Washington using an online testing format. The sample in this study took the MSP as part of that online cohort. Students were tested within their language arts class for the reading assessment and math class for the mathematics portion of the assessment. The reading portion of the MSP asks students to read a passage and select from a series of multiple choice options. In addition, readers respond in writing with short and extended response questions. The test period was approximately two hours long, but the students were given options to continue testing for as long as they needed. The 40 items on this assessment consist of multiple choice, short answer and completion questions. This assessment used norm and criterion scoring to evaluate individual student progress. As 2010 was the first year of the assessment's implementation, reliability and validity scores were not established. Rasch modeling was used to create scaled scores provided by the State of Washington as follows.

Level 4- Above Standard represents superior performance.

Level 3- Meets Standard represents solid academic performance.

Level 2- Below Standard represents partial knowledge and performance of academic skill

Level 1- Well Below Standard represents little or no knowledge or skill in an area.

For this study, data from the Spring 2010 MSP as well as the Spring 2011 MSP were used. In this analysis, 2010 MSP acted as a predictor variable and 2011 MSP served as the

outcome measure. The MSP, as an outcome measure, serves primarily as dichotomous data in practice. That is, the significance of the assessment is whether a student meets benchmark or not. For that reason, I coded the data for the 2011 outcome measure. Students who obtain a score of 400 or greater were assigned a score of one, while students who score below 400 were assigned a score of zero. This will assist me in determining the predictive value each screening tool has on the outcome measure.

The second measure of Reading Comprehension is *The Gates-MacGinitie Reading Test* (GMRT) of Comprehension. The reading comprehension section of the GMRT consists of reading passages written in a variety of writing styles related to various content areas. The student's task is to choose the correct answers which will demonstrate both literal and inferential comprehension of questions about each passage. The assessments are group administered. The entire assessment, vocabulary and comprehension, takes 55 minutes to administer.

Classification accuracy, reliability, validity, and generalizability statistics are reported by the National Center on Response to Intervention for Vocabulary and Comprehension assessments combined into a composite score. However, these results are not reported for sixth grade. The closest grade level with data analysis is fourth grade. Classification accuracy rates for the composite score (vocabulary and comprehension) for fourth grade students are: sensitivity .73 and specificity of .88, and an overall classification accuracy rate of .91. Reliability is reported as KR-20 for fourth grade is .96. In addition, test-retest median score is measured at .90. Criterion and construct validity measures for fourth grade are .92 and .80 respectively. When measured for multiple grade levels (two, four, six, eight and ten) construct validity was measured at .90. Generalizability is limited for this assessment as the studies are limited to one state

primarily with English speaking general education students in the far western portion of the United States.

Oral Reading Fluency. AIMSweb passages and norms were used to measure oral reading fluency (ORF; Shinn & Shinn, 2002). It is an individually administered assessment based on listening to students read grade-level passages aloud for one minute, and calculating the number of words read correctly (Shinn & Shinn, 2002). Hence, the score is the number of words correct per minute (wcpm).

Classification accuracy, reliability, validity, and generalizability statistics are reported by the National Center on Response to Intervention (2010) for sixth grade AIMSweb (Howe & Shinn, 2002) Reading Assessments. AIMSweb introduced National Norms in the fall of 2011 to provide norms that reflect the national student population from grades 1 through 8 on ORF, with mid-interval scoring method used for calculating local and national percentiles. They found that, at sixth grade, the AIMSweb winter passage ORF norm is $M = 154$ ($SD = 40$). Classification accuracy for sixth grade students at the fall assessment period showed a sensitivity of .77 and a specificity of .74, with overall classification accuracy rate of .75. When sensitivity is set at .90, which is the recommended level for sensitivity measures (Jenkins, 2007), specificity drops to .59. During the winter assessment window, sensitivity was.78 and specificity was .73, with an overall classification accuracy of .74. With sensitivity set at .90, specificity was recorded at .58. The similarity between the scores across measurement periods would seem to indicate that the time lapse between the screening tool and the outcome measure, in this case the Illinois Standards Achievement Test (ISAT), was irrelevant. Median sixth grade reliabilities were: .97 for alternate-form reliability, .99 for inter-scorer reliability, .96 for split-half reliability, and .95 for test-retest reliability. Median predictive construct validity for sixth grade using the Illinois

Standards Achievement Test (ISAT) as the criterion measure averaged .64. Finally, Christ and Silberglitt (2007) evaluated benchmark data for 8,200 students in grades 1 through 5 over the course of eight consecutive school years, with AIMSweb passages were used in three of those years. Correlations between benchmark scores for adjacent seasons (fall-winter or winter-spring) were measured at each grade level. In fifth grade, the fall-winter correlation was 0.92 and the winter-spring correlation was 0.93.

Vocabulary. Vocabulary as a stand-alone construct was measured with a curriculum-based measure (CBM). Even though the Gates-MacGinitie Vocabulary test (discussed subsequently) is group administered, it is rather time-consuming with a combined (Vocabulary and Comprehension subtests) testing time of 55 minutes. For this reason, a vocabulary curriculum-based measure (CBM) was employed by the school district and included in this study to determine whether it might be used as a potential substitute/proxy measure for predicting future reading achievement performance. This CBM was developed using words from the school's science and social studies curriculum. Twenty-five vocabulary words are available in the "word bank" at the top of the assessment, while 20 definitions make up the questions. Students are asked to select the most appropriate word for the definition, and write it in the space provided. Five distracter words remain unused from the word bank to help reduce the impact of guessing. The items selected are academic in nature and are less likely to be obtained in general conversation. Items were selected to represent a range of difficulty. In creating the test, items were selected across a wide continuum of difficulty to ensure that few students would obtain perfect or zero scores, thereby minimizing floor and ceiling effects. The bulk of the items were targeted at the intermediate to low level, as the students who fall in this range are tend to be the ones for whom the most information is required to make screening determinations. This

assessment was limited to 20 minutes and was given at the beginning of the sixth grade school year.

Silent Reading Fluency. The TOSRE (Vocab + Comp) is a group-administered assessment that measures a student's contextual reading ability. It requires students to silently read sentences of increasing difficulty, and to make a judgment regarding the accuracy of the sentence with a response of yes of no. Sentence length ranges from four to ten words. Examples of these sentences are: "You will always find icebergs at the beach" or, "The sky is above the trees". This assessment measures both silent reading fluency and a simple form of comprehension. Students are given exactly three minutes for this assessment. The scoring procedures correct for the dichotomous nature of the assessment and student guessing, by grading the assessment as the number of correct answers minus the number of errors. According to the publisher, the average alternate-form reliabilities across grade levels (first through twelfth) for immediate administration range from .86 to .95, and the averages for alternate-form delayed administration range from .86 through .95. The inter-scorer reliabilities were .99.

In a study conducted by Denton et al. (2011), in grades six and eight, the TOSRE (Vocab + Comp) was more strongly related to reading comprehension as measured by the Texas Assessment of Knowledge and Skills (TAKS) with a correlation coefficient of .56 compared to ORF .50 or Maze .40. Regardless of the benchmark selected, TOSRE (Vocab + Comp) outperformed ORF in overall classification accuracy: 25th percentile ORF .58, TOSRE (Vocab + Comp) .62; 50th percentile ORF .66, TOSRE (Vocab + Comp) .69; 75th percentile ORF .63, TOSRE (Vocab + Comp) .67.

Vocabulary + Comprehension Composite. The group-administered Gates-MacGinitie (Vocab + Comp) composite includes comprehension and vocabulary subtests. (For fourth grade

and beyond, the Vocabulary and Comprehension assessments can be administered at the same time.) The Comprehension measure consists of reading passages written in a variety of writing styles related to various content areas. The student's task is to choose the correct answers which will demonstrate both literal and inferential comprehension of questions about each passage. The Vocabulary subtest is divided by levels for grades 3 through 12 and measures the student's ability to choose the word or phrase that is closest in meaning to the test word. The context is intended to suggest the part of speech but not to provide clues as to the meaning of the word. In this way, the assessment limits the amount of influence the Context processor has in this measure. (Importantly, the Vocabulary assessment measures *reading*, rather than *oral*, vocabulary. This differentiation should not go unnoticed as the assessment results for the Gates-MacGinitie (Vocab + Comp) are dependent on vocabulary knowledge and foundational reading skills. In an oral vocabulary test, such as the *Peabody Picture Vocabulary Test* (PPVT Dunn & Dunn, 1997), the assessment isolates the skill of vocabulary knowledge from that of decoding and reading fluency.) The Gates-MacGinitie (Vocab + Comp) subtests are norm-referenced achievement tests designed to provide a measure of overall reading achievement. The entire assessment, Vocabulary and Comprehension together, takes 55 minutes to administer.

Classification accuracy, reliability, validity, and generalizability statistics are reported by the National Center on Response to Intervention for Vocabulary and Comprehension assessments combined into a composite score. However, these results are not reported for sixth grade. The closest grade level with data analysis is fourth grade. Classification accuracy rates for the composite score (vocabulary and comprehension) for fourth grade students are: sensitivity .73 and specificity of .88, and an overall classification accuracy rate of .91. Reliability is reported as KR-20 for fourth grade is .96. In addition, test-retest median score is measured at .90.

Criterion and construct validity measures for fourth grade are .92 and .80 respectively. When measured for multiple grade levels (two, four, six, eight and ten) construct validity was measured at .90. Generalizability is limited for this assessment as the studies are limited to one state primarily with English speaking general education students in the far western portion of the United States.

Data Analysis

Data analyses were conducted in concert with each research question. Since these data are from one middle school, I treated students as independent of each other for purposes of statistical analyses (i.e., not nested within organizational hierarchies, like classrooms). This was a reasonable assumption because most students have six or seven classes within a given semester, and classes can change from one semester to the next. As such, ignoring multiple classroom memberships would not appear to violate the independence assumption.

Specific variables used for analyses are detailed in the previous section. Recall that the focus of this study is whether or not a child (sixth-grader) passed the 2011 Reading MSP, a high-stakes test in Washington State. Although dichotomization of a scale score is generally not recommended for statistical analyses (i.e., the loss of information can lead to a severe loss in power), the pass/fail cut is an important policy outcome that should be examined in the context of my wish to determine the psychometric ability of reading screening measures to accurately predict who passes and who does not.

To reiterate, the following research questions and data analyses were as follows.

1. What are the direct, one-to-one relationships among 2010 demographic, prior MSP, and reading screening measures with 2011 MSP Reading passing status? In other words, do we see significant relationships among all measures? What is the strength of those relationships?

- Bivariate (zero-order) correlations were conducted.

2. What are the area-under-the-curve (AUC) values for each individual reading screening measure predicting 2011 MSP Reading passing status (without adjustment for each other, the prior year MSP score, or demographic characteristics)?

- The diagnostic performance of a given measure to discriminate "at risk" cases from normal cases will be evaluated using Receiver Operating Characteristic (ROC) curve analysis (Metz, 1978; Zweig & Campbell, 1993). For this study, classification accuracy as measured by AUC (sensitivity and specificity) was examined descriptively.

3. What are the unique contributions of student demographic characteristics and prior year MSP Reading score to predicting 2011 MSP Reading passing status (henceforth "baseline covariates")?

- This question was answered in Block 1 of a multiple logistic regression analysis, with sequential predictor entry. Block 1 of this model includes the following predictors:

  Log Odds(2011 MSP Passing) = Prior MSP(2010) + Female + SPED + ELL + FRL.

4. What is the contribution of each individual reading measure to predicting 2011 MSP Reading passing status, after adjustment for (above and beyond) baseline covariates?

- This question was tested using multiple logistic regressions, with Block 1 as specified above and Block 2 specified with each individual reading screening measure separately, as follows:

  Log Odds(2011 MSP Passing) = Block1 + ORF

  Log Odds(2011 MSP Passing) = Block1 + CBM Vocab

  Log Odds(2011 MSP Passing) = Block1 + TOSRE (Vocab + Comp)

Log Odds(2011 MSP Passing) = Block1 + Gates MacGinitie (Vocab + Comp)

5. What are the contributions of the combined set of reading measures to predicting 2011 MSP Reading passing status, after adjustment for (above and beyond) baseline covariates?

- Similar to above, this question was also evaluated using multiple logistic regression, but this time Block 2 included all reading screening measures together.

Log Odds(2011 MSP Passing) = Block1 + ORF + CBM Vocab + TOSRE (Vocab + Comp) + Gates-MacGinitie (Vocab + Comp)

6. What are the individual and combined contributions of all reading measures to 2011 MSP Reading passing status, without prior year 2010 MSP Reading scale score used as a predictor? In other words, what does the model look like for children who might not have been assessed on the prior year test, such as students who move frequently?

- This question was assessed using multiple logistic regression with sequential predictor entry just like research questions 4 and 5, but this time Fall 2010 MSP was not be used as a predictor.

In all of the logistic regression models, I computed predicted probabilities for passing the 2011 MSP to provide a more direct interpretation of the estimated effects (other than in "logits").

CHAPTER FOUR
Results

Results from data analysis are reported in this chapter. Descriptive statistics are reported first and then each analysis in order of research questions, including zero-order correlations, ROC curve analyses, and multiple logistic regression analyses.

Descriptive Statistics

Descriptive statistics were calculated for demographic characteristics (gender, special education (SPED) status, English language learner (ELL) status, and free and reduced lunch (FRL) status), as well as all of the reading variables, including: Measurement of Student Progress (MSP), AIMSweb oral reading fluency (ORF), curriculum-based measure of vocabulary (CBM Vocab Test of Sentence Reading Efficiency Vocabulary and Comprehension score (TOSRE (Vocab + Comp)), and Gates-MacGinitie (Vocab + Comp). These data are presented in Tables 2 (demographic variables) and 3 (reading measure variables).

Table 2.

Sample Demographic Characteristics

Variable	*Percent*
Gender	
Male	56.20%
Female	43.80%
SPED	
Eligible	12.70%
Ineligible	87.30%
ELL	
Eligible	2.30%
Ineligible	97.70%
FRL	
Eligible	55.90%
Ineligible	44.10%

Table 3.

Descriptives for all Reading Measures

Measure	*M (SD)*	*Range*
Reading MSP		
Spring 2010 Grade 5	408.31 (28.45)	320 - 475
Spring 2011 Grade 6	405.53 (27.59)	330 - 471
ORF		
Fall 2010 Grade 6	148.51 (45.94)	24 - 276
CBM Vocab		
Fall 2010 Grade 6	11.21 (4.65)	0 - 20
TOSRE (Vocab + Comp)		
Fall 2010 Grade 6	26.82 (11.51)	1 - 55
Gates MacGinitie (Vocab + Comp)		
Fall 2010 Grade 6	50.55 (21.00)	0 - 91

As mentioned previously, the sixth grade (Spring 2011) MSP reading subtest outcome variable was dichotomized into passing status, where "pass" = scores ≥ 400, and "not pass" = scores < 400 (as per the state criterion). After dichotomization, it was observed that 67% of the students in the sample passed the state test, and 33% did not. All subsequent analyses employ 2011 MSP passing status as the focal outcome of interest.

Correlations among Variables

The first research question in this study asked: What are the direct, one-to-one relationships among 2010 demographic, prior MSP, and reading screening measures with 2011 MSP Reading passing status? Bivariate (zero-order) correlations, using all variables across the sample, were conducted to answer this question. The results are presented in Table 4. While most of the demographic variables were correlated with the dependent variable (MSP 2011 passing status), it was notable that FRL negatively correlated ($p < .001$) with MSP in 2010, but was not significantly correlated with MSP in 2011; nevertheless, the direction of the relationship was in the same direction (just a weakened relationship by 2011). FRL, SPED, and ELL status each

correlated negatively with all reading screening measures (*p*s < .001) The strongest positive correlates with the dependent variable, MSP 2011 passing status, was prior year 2010 MSP and the Gates-MacGinitie (Vocab + Comp) screening measure (*r*s = 0.64, *p*s < .001). In addition, each of the other reading measure predictors (ORF, CBM Vocab, and TOSRE Vocab + Comp) had direct positive relationships averaging *r* = 0.47 with the outcome, as well as significant positive correlations with each other and the prior year MSP (*r*s ranged from .57 to .81, *p*s < .001). Notably, the reading measures' predictive validity with the outcome appeared to be lower than concurrent validity among one another.

Area Under the Curve (AUC) Results

The second research question in this study examines the question: What is the area-under-the-curve (AUC) value for each individual reading screening measure with 2011 MSP Reading passing status, without adjustment for the prior year MSP score or demographic characteristics? The diagnostic performance of a given measure to discriminate "at risk" cases from normal cases was evaluated using Receiver Operating Characteristic (ROC) curve analysis (Metz, 1978; Zweig & Campbell, 1993). An AUC greater that .90 is considered excellent; .80 to.90, good; .70 to .80, fair; and below .70, poor (Compton, Fuchs, Fuchs & Bryant, 2006). Results of this analysis indicated that prior year 2010 MSP and the Gates-MacGinitie (Vocab + Comp) measures were most predictive (AUCs = 0.89), with CBM Vocab in second place with AUC = 0.81; all three are considered to have "good" predictive value. The areas for the ORF and TOSRE Vocab + Comp measures were, on the other hand, considered "fair" predictors (AUCs = 0.76 and 0.79, respectively). See Figure 5 for an illustration of each curve.

Table 4.

Zero-order Correlations among the Outcomes and Predictors

Measure	*M*	*(SD)*	1.	2.	3.	4.	5.	6.	7.	8.	9.
Outcome											
1. Spring 2011 Reading MSP (1=pass)	0.67	(0.47)	--								
Baseline Predictors (Spring 2010)											
2. MSP Scale Score	408.31	(28.45)	.64 ***	--							
3. Female (1=yes, 0=no)	0.44	(0.50)	.17 **	.17 **	--						
4. FRL (1=yes, 0=no)	0.56	(0.50)	-.11	-.25 ***	.04	--					
5. SPED (1=yes, 0=no)	0.13	(0.33)	-.40 ***	-.49 ***	-.16 **	.08					
6. ELL (1=yes, 0=no)	0.02	(0.15)	-.13 *	-.17 **	.04	.14 *	-.06				
Reading Measure Predictors (Fall 2010)								--			
7. ORF	148.51	(45.94)	.45 ***	.65 ***	.16 **	-.21 ***	-.49 ***	-.09	--		
8. CBM Vocab	11.21	(4.65)	.52 ***	.71 ***	.03	-.25 ***	-.45 ***	-.19 ***	.57 ***	--	
9. TOSRE (Vocab + Comp)	26.82	(11.51)	.47 ***	.68 ***	.11	-.24 ***	-.48 ***	-.21 ***	.77 ***	.63 ***	--
10. Gates MacGinitie (Vocab + Comp)	50.55	(21.00)	.64 ***	.81 ***	.14 *	-.26 ***	-.51 ***	-.16 **	.68 ***	.71 ***	.72 ***

Note. N = 306 sixth grade students from one public middle school. MSP= Measurement of Student Progress; FRL=Free and Reduced Lunch; SPED= Special Education; ELL= English Language Learner; ORF= Oral Reading Fluency; CBM= Curriculum Based Measurement; TOSRE (Vocab + Comp)= Test of Sentence Reading Efficiency.

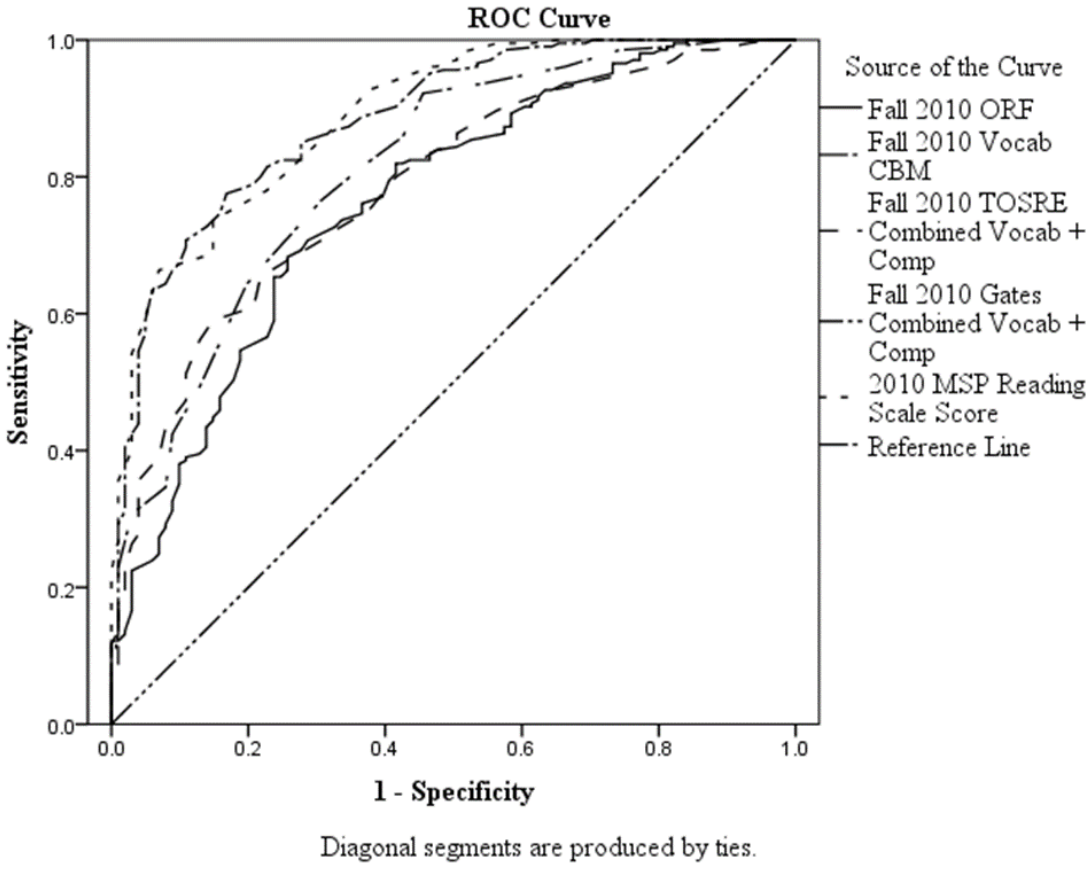

Figure 5. *ROC curve results for reading predictor variables*

Predictors of Reading Achievement

The third research question: What are the unique contributions of student demographic characteristics and prior year MSP Reading score to predicting 2011 MSP Reading passing status (henceforth "baseline covariates")?**,** was analyzed by conducting a multiple logistic regression on MSP Reading passing status using a set of baseline covariates, including: prior year MSP, gender (Female status), special education (SPED) status, English Language Learner (ELL) status, and Free and Reduced Lunch (FRL) status). For ease of results interpretation, all demographic variables were effect coded (+1=yes, -1=otherwise), and prior year MSP was standardized into z-scores. The overall model results (see Table 5 upper portion) showed that the set of predictors reliably distinguished individuals who passed the MSP from those who did not, χ^2 (5) =166.71, *p*

< .001. The approximate variance in MSP passing status that was accounted for by the set of predictors was .58 using Nagelkerke's formula. Model sensitivity was 92% and specificity was 65%, with an overall hit rate of 83%, which was better than the null model's hit rate of 67%. Importantly, although we observed direct (one-to-one) relationships among the baseline covariates and the outcome (see again Table 4), results for this regression model also showed that only prior year MSP was *uniquely* predictive of 2011 MSP passing status. One way to interpret this finding is that, for every standard deviation increase in prior year MSP, there is a predicted 2.57-logit increase in MSP passing, holding all other variables constant. Another way to interpret this finding is that students with relatively higher prior year MSP (+1 *SD*) had a 97% probability of passing the current MSP compared to students scoring at average the prior year (70%) or students with relatively low prior year MSP (-1 *SD*) who had only an 18% probability of passing the current year MSP (see Figure 6 for plot of predicted probabilities).

Table 5.
Multiple Logistic Regression Model Results

	χ^2 (df)	*p*	*Pseudo* R^2	*Sens*	*Spec*	*HR*	*b*	*(SE)*	*Wald* (1)	*p*
Baseline Block	166.71(5)	<.001	0.58	0.92	0.65	0.83				
Intercept							0.86	(0.61)	1.99	.158
Reading MSP							2.57	(0.36)	52.18	<.001
Female							0.16	(0.17)	0.84	.360
SPED							-0.28	(0.27)	1.02	.312
ELL							-0.19	(0.53)	0.13	.721
FRL							0.10	(0.18)	0.33	.568
Individual Reading Measures (Block 2)										
ORF	0.20(1)	.654	0.59	0.90	0.65	0.82	0.12	(0.26)	0.20	.655
CBM Vocab	2.08(1)	.149	0.59	0.89	0.63	0.80	0.34	(0.24)	2.03	.154
TOSRE (Vocab + Comp)	0.15(1)	.697	0.59	0.90	0.65	0.82	0.10	(0.25)	0.15	.697
Gates MacGinitie (Vocab + Comp)	12.71(1)	<.001	0.62	0.91	0.67	0.83	1.03	(0.30)	11.68	.001
Combined Reading Measures (Block 2)	14.01(4)	.007	0.62	0.90	0.68	0.83				
ORF							-0.92	(0.33)	0.08	.778
CBM Vocab							0.13	(0.26)	0.23	.633
TOSRE (Vocab + Comp)							-0.26	(0.32)	0.63	.426
Gates MacGinitie (Vocab + Comp)							1.12	(0.34)	11.01	.001

Note. N = 306 sixth grade students from one public middle school. MSP= Measurement of Student Progress; FRL=Free and Reduced Lunch; SPED= Special Education; ELL= English Language Learner; ORF= Oral Reading Fluency; CBM= Curriculum Based Measurement; TOSRE (Vocab + Comp) = Test of Sentence Reading Efficiency. The pseudo-R-squared value reported is the approximate total variance accounted for, as calculated by Nagelkerke. All status variables were effect-coded and reading measure variables standardized (in z-scores for analyses for ease of results interpretation).

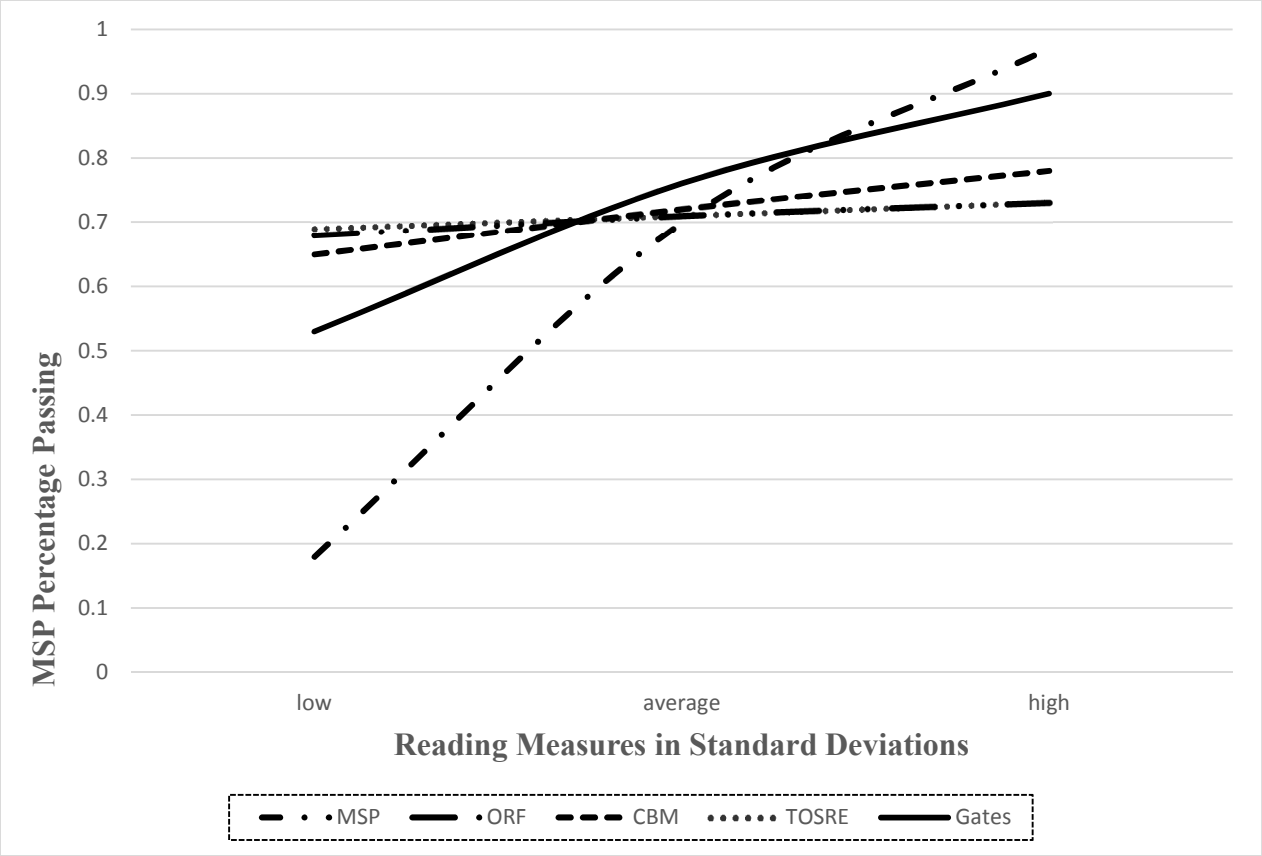

Figure 6. Predicted probabilities with prior year MSP

Predicting Reading Achievement with Prior Year Reading Achievement

The fourth research question, "What is the contribution of each individual reading measure to predicting 2011 MSP Reading passing status, after adjustment for (above and beyond) baseline covariates?", was analyzed using multiple logistic regressions, with Block 1 as specified above and Block 2 with each individual reading screening measure, as follows:

Log Odds (2011 MSP Passing) = base + ORF

Log Odds (2011 MSP Passing) = base + CBM Vocab

Log Odds (2011 MSP Passing) = base + TOSRE (Vocab + Comp)

Log Odds (2011 MSP Passing) = base + Gates-MacGinitie (Vocab + Comp)

The results of these analyses are also presented in Table 5 (middle portion), and indicate that ORF (p = .66), CBM Vocab (p = .15), and TOSRE Vocab + Comp (p = .70) were not uniquely predictive of MSP passing status when the base model covariates (Prior year MSP + demographics) were controlled for. Overall classification accuracies were 82%, 80%, and 82%, respectively, which wasn't better than the base model classification accuracy of 83%. In contrast, the Gates-MacGinitie (Vocab + Comp) was uniquely predictive of passing status (and showed a classification accuracy of 83% equivalent to the base model). One way to interpret the slope coefficient findings is that, for every standard deviation increase in Gates-MacGinitie (Vocab + Comp), there is a predicted 1.03-logit increase in MSP 2011 passing status, holding the base model variables constant. Another way to interpret the meaning of the findings is with predicted probabilities: students with relatively higher Gates-MacGinitie (Vocab + Comp) scores had a 90% probability of passing the current MSP compared to only 53% probability for students with lower Gates-MacGinitie (Vocab + Comp) scores (see Figure 6).

The final model tested in this series was one with all predictors entered into the model (Table 5, third portion). With all reading predictors in the model, the approximate variance in MSP passing status accounted for by the set of predictors was .62 using Nagelkerke's formula, and overall classification accuracy was 83% (same as with just the Gates-MacGinitie (Vocab + Comp) alone).

Predicting Reading Achievement without Prior Year Reading Achievement

The sixth research question posed in this study was: What are the individual and combined contributions of all reading measures to 2011 MSP Reading passing status, without prior year 2010 MSP Reading scale score as a covariate? In other words, what does the model look like for children who might not have been assessed on the prior year's test, such as students

who move frequently? This question was assessed again using multiple logistic regression with sequential predictor entry (just like questions 4 and 5), but this time Fall 2010 MSP was omitted from analysis. The results of this analysis are presented in Table 6. The base model was again significant, but the approximate variance in MSP passing status accounted for by the set of predictors was only .25 using Nagelkerke's formula (compare with the previous base model results in Table 5, where the proportion explained was .58). Model sensitivity was 96% and specificity was 38%, with an overall hit rate of 77%, which was better than the null model's hit rate of 67% (but a weaker performance compared to the base model with the prior year MSP included, which had a classification accuracy of 83%).

Interestingly, in the base model without prior year MSP, Female status, SPED status, and ELL Status, were each uniquely predictive of MSP passing status while FRL status was not. Females had a 0.62-logit advantage over males (double the coefficient when effect coding is used, as it was here), whereas students designated SPED had a 2.56-logit deficit compared to peers without the designation, and students designated ELL had a 2.02-logit deficit compared to non-ELL students. In other words, females had an 31% predicted probability of passing whereas males' predicted probability was 20%; SPED and ELL students had 9% and 11% predicted probabilities of passing, respectively.

When prior year MSP is not available, *all* of the reading measures selected for this study add predictive power beyond baseline demographic information (see Table 6 middle portion for results, as well as Figure 7 for illustration of predicted probabilities by measure). For every standard deviation increase on ORF, there is a predicted 0.94-logit increase in MSP passing, holding all other variables constant (students who scored relatively higher had a 63% probability of passing compared with 20% for lower performers).The overall classification accuracy was

77% with this measure added to the baseline demographics (which is the same accuracy as baseline alone).Similarly, for every standard deviation increase on the CBM Vocab measure, we expect a 1.27-logit increase in MSP passing status, holding all other variables constant (there was a 78% probability of passing the MSP for students with relatively high vocabulary scores, compared with 22% for lower performers). The overall classification accuracy with this measure added to baseline was slightly better than baseline alone, at 79% compared with 77%. The TOSRE Vocab + Comp also showed the similar predictive utility as the ORF and CBM Vocab measures, with a 76% correct classification rate, and a 72% predicted probability of passing for higher performers and a 26% probability for lower performers).

Of importance, the Gates-MacGinitie (Vocab + Comp) measure had the best predictive utility, with an overall classification accuracy rate of 83% (an improvement of 6% over baseline covariates). As Figure 7 illustrates, students who score higher on this measure had a 93% probability of passing the MSP compared with a 21% probability of passing for lower-performing students.

Finally, as shown in Table 6 (bottom portion), if all reading predictors are added together to the base model with demographic information only, we find the classification accuracy to be the same as with Gates-MacGinitie (Vocab + Comp) alone (83%); however, we also find that both CBM Vocab and Gates-MacGinitie (Vocab + Comp) each *uniquely* contribute to the model ($ps < .05$). Hence, the CBM Vocab does offer some predictive utility above and beyond the effects of the demographics and the Gates-MacGinitie (Vocab + Comp).

Table 6.
Multiple Logistic Regression Model Results No Prior Year Assessment Available

	χ^2 (df)	*p*	*Pseudo* R^2	*Sens*	*Spec*	*HR*	*b*	*(SE)*	*Wald* (1)	*p*
Baseline Block	60.34(4)	<.001	0.25	0.96	0.38	0.77				
Intercept							-1.10	(0.48)	5.21	.023
Female							0.31	(0.14)	4.80	.028
SPED							-1.28	(0.22)	32.35	<.001
ELL							-1.01	(0.43)	5.41	.020
FRL							1.15	(0.14)	1.19	.276
Individual Reading Measures (Block 2)										
ORF	26.67(1)	<.001	0.34	0.92	0.45	0.77	0.94	(0.20)	22.34	<.001
CBM Vocab	51.35(1)	<.001	0.43	0.90	0.56	0.79	1.27	(0.20)	40.15	<.001
TOSRE (Vocab + Comp)	28.86(1)	<.001	0.35	0.91	0.44	0.76	0.96	(0.20)	24.40	<.001
Gates MacGinitie (Vocab + Comp)	93.62(1)	<.001	0.55	0.91	0.67	0.83	1.94	(0.25)	61.41	<.001
Combined Reading Measures (Block 2)	97.78(4)	<.001	0.56	0.90	0.68	0.83				
ORF							0.05	(0.30)	0.03	.858
CBM Vocab							0.49	(0.25)	3.95	.047
TOSRE (Vocab + Comp)							-0.08	(0.30)	0.07	.786
Gates MacGinitie (Vocab + Comp)							1.68	(0.30)	30.64	<.001

Note. N = 306 sixth grade students from one public middle school. MSP= Measurement of Student Progress; FRL=Free and Reduced Lunch; SPED= Special Education; ELL= English Language Learner; ORF= Oral Reading Fluency; CBM= Curriculum Based Measurement; TOSRE= Test of Sentence Reading Efficiency. The pseudo-R-squared value reported is the approximate total variance accounted for, as calculated by Nagelkerke. All status variables were effect-coded and reading measure variables standardized (in z-scores for analyses for ease of results interpretation).

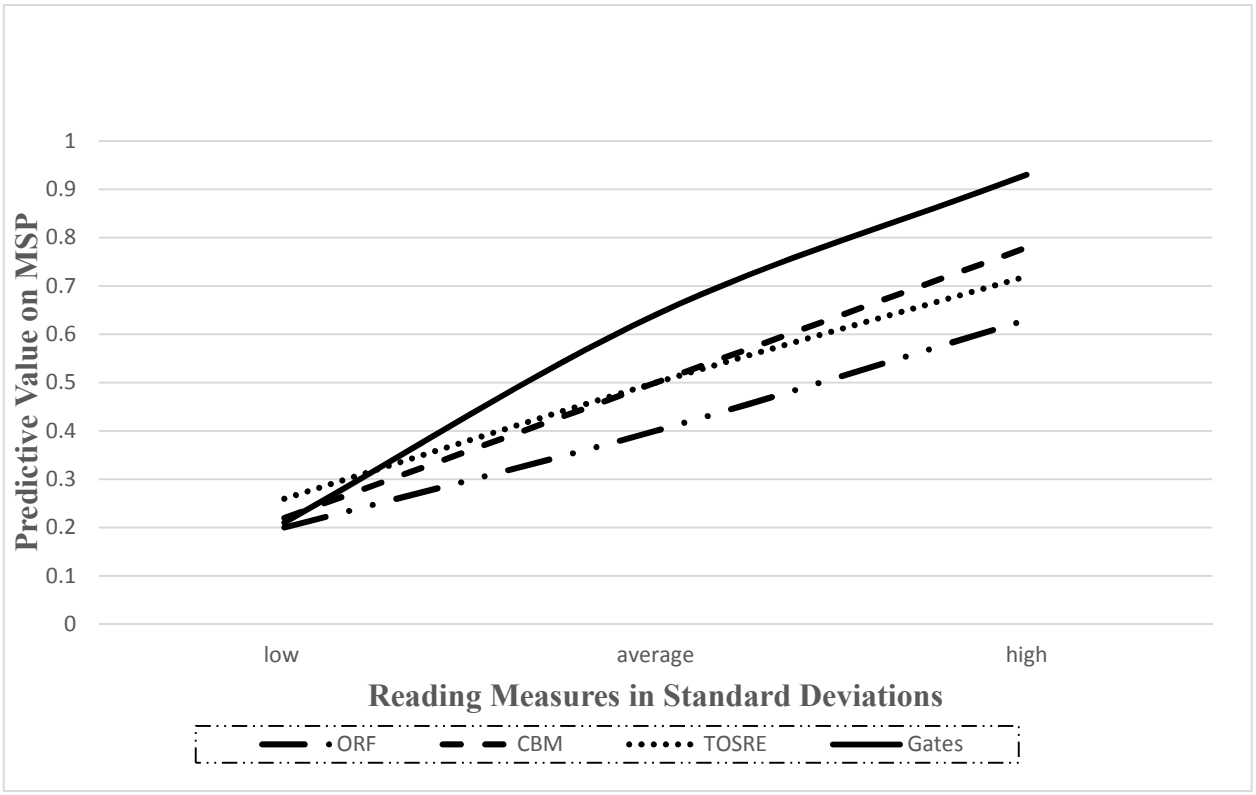

Figure 7. Predicted probabilities without prior year MSP

Optimal Set of Screening Assessments and Information for Predicting Reading Achievement? Two Exploratory Models

As described earlier, the quality of screening tools are evaluated using three primary factors: accuracy, efficiency and consequential validity. With these characteristics and the prior results in mind, two additional models were created and tested as an exploration of optimal combinations of predictors. For each of these models, prior year MSP and Gates-MacGinitie (Vocab + Comp) were excluded, as both were already shown to account for a large portion of the variance in the initial modeling and to have good predictive utility when combined with demographics (83% correct classifications). Further, both are extremely time-consuming to administer and require extensive training and the MSP score may not always be available.

Instead, two additional models were estimated to address whether an efficiency-oriented set of variables (ORF and TOSRE Vocab + Comp) and/or a consequential validity-oriented set of variables (ORF and CBM Vocab) would demonstrate predictive utility. In each of these models, baseline demographic covariates were retained as Block 1.

Efficiency-Oriented Model. Distal measures of reading comprehension, like the ORF and the TOSRE Vocab + Comp, are quicker to administer than assessments such as the Gates-MacGinitie (Vocab + Comp), using approximately 83% less instructional time for testing. Findings from this model (see Table 7) showed that the combination did reliably distinguish individuals who passed the MSP from those who did not.

Consequential Validity-Oriented Model. Because false-negatives are associated with issues in consequential validity, we theorized adding a measure to ORF may help control for this problem. To this end, we referred back to our conceptual framework to guide selection of this measure. ORF (a distal measure of fluency) and CBM Vocab (an untimed measure of written vocabulary) are part of separate constructs. For this reason, it is possible that the CBM Vocab might combine with ORF to assist in improved classification accuracy. In addition, examination of the correlation table shows that, while ORF and CBM Vocab are correlated with one another (see Table 4; $r = 0.57$) they share the lowest relationship between ORF and other measures. The results of this model (lower portion of Table 7) show that the approximate variance in MSP passing status accounted for by this model was .42 with an overall hit rate of 80%. By combining ORF with CBM Vocab, sensitivity remained above 90% and specificity, while still low, was increased by 38% (base model alone) to 58%. In fact, this model shows an overall classification accuracy improvement of 3% to 80%, which, while not perfect, may be useful to schools who lack the resources to administer more time-consuming and costly tests.

Table 7.

Multiple Logistic Regression Composite Models Results No Prior Year Assessment Available

	χ^2 (df)	*p*	*Pseudo* R^2	*Sens*	*Spec*	*HR*	*b*	*(SE)*	*Wald* (1)	*p*
Baseline Block	60.34(4)	<.001	0.25	0.96	0.38	0.77				
Intercept							-1.10	(0.48)	5.21	.023
Female							0.31	(0.14)	4.80	.028
SPED							-1.28	(0.22)	32.35	<.001
ELL							-1.01	(0.43)	5.41	.020
FRL							1.15	(0.14)	1.19	.276
Exploatory Combined Models Block 2										
Efficiency Model	33.57(2)	<.001	0.37	0.90	0.47	0.76				
ORF							0.53	(0.25)	4.53	.033
TOSRE							0.63	(0.25)	6.58	.010
Consequential Validity Model	58.63(2)	<.001	0.45	0.91	0.58	0.80				
CBM Vocab							1.09	(0.21)	27.06	<.001
ORF							0.58	(0.22)	6.90	.009

Note. N = 306 sixth grade students from one public middle school. MSP= Measurement of Student Progress; FRL=Free and Reduced Lunch; SPED= Special Education; ELL= English Language Learner; ORF= Oral Reading Fluency; CBM= Curriculum Based Measurement; TOSRE= Test of Sentence Reading Efficiency. The pseudo-R-squared value reported is the approximate total variance accounted for, as calculated by Nagelkerke. All status variables were effect-coded and reading measure variables standardized (in z-scores for analyses for ease of results interpretation).

CHAPTER FIVE

Discussion

Overall, the purpose of this study was two-fold: first, to evaluate the effectiveness of universal screening tools for students in sixth grade, then to determine if a combination of screening measures might increase the classification accuracy rate of these tools. It was predicted that a combination of screeners would reduce the number of false negative results and therein improve consequential validity of the process. Specifically, this research sought to: (1) evaluate the one-to-one relationships among 2010 demographic, prior MSP and reading screening measures with 2011 MSP; (2) compare AUC values for each individual screening measures with 2011 MSP; (3) identify the unique contributions of extant data (student demographic characteristics and prior year MSP) in predicting 2011 MSP results; (4) identify the unique contributions of each individual screening measure (ORF, CBM vocabulary, TOSRE (Vocab + Comp), and Gates-MacGinitie (Vocab + Comp)) in predicting 2011 MSP results; (5) identify the contributions of the combined set of screening measures when predicting 2011 MSP results; and finally, (6) when prior year MSP is excluded from the model, what are the individual and potential combinations of universal screening tools that predict 2011 MSP.

Limitations

As with all studies, there are some limitations to the findings presented here. Despite these limitations, the results provide an important window into assessment for schools using an RTI framework with middle school students. Simply translating what has been established in elementary grades to students in middle school is not appropriate.

First, the extant data for this study came from one school district at a single middle school with only 6th grade students. The data was collected from the state of Washington,

therefore generalizing to other states is limited. In addition, the researchers did not collect the data and therefore do not have full knowledge regarding the assessment procedures. Finally, although the data provided came from a relatively diverse population, the percentage of ELL students was well below state averages.

Summary and Interpretation of Findings

The findings of this study provide similar psychometric results reported from prior research (Jenkins et al., 2007) and extend the findings into 6th grade. It suggests that, while correlations between universal screening measures and an outcome measure are statistically significant, classification accuracy results identify potential problems for application in school settings. Specifically, this study indicates that proximal measures of reading comprehension, such as prior year MSP and the Gates-MacGinitie (Vocab + Comp), outperform distal measures, like ORF and TOSRE (Vocab + Comp), for predicting the reading achievement of students in sixth grade.

Descriptive Statistics

The demographic descriptive statistics from this **st**udy represent a school that closely mirrors state statistics on several components. Data from this study were reported from the only middle school in a small suburban school district in the Pacific Northwest. The school reported a minority population of 39% compared to the state average of 42%; special education closely mirrored the state with 13% and 13.2% respectively; and 57% of the students at this school received free or reduced lunch compared to a state average of 46%. Because of this, I was surprising to see that, despite the diversity reported by the school, students who were identified as ELL (3%) was notably lower than the state average (nearly 10%). This disproportionality is important to note as we evaluate the data. In contrast to the demographics of this study, Medina

(2012), evaluated reading factors that influenced performance on the FCAT where 96% of the student population was Hispanic. Medina found that vocabulary was a strong predictor of FCAT performance amongst eleventh grade students. Similarly, Valencia et al., 2010 found that the addition of passage comprehension to a three-factor model with rate, accuracy, and prosody accounted for significantly more variance on the outcome measure of comprehension. Valencia & colleagues' (2010) research was conducted in a school where one third of the students were classified as ELL; data from this study only contained three percent of the students identified as ELLs. Our findings help support the need for more direct measures of reading comprehension regardless of the percentage of ELL students.

Of particular interest in the descriptive statistics results for the reading measures is the relatively high base rate of 33% for the outcome measure (MSP 2011). This base rate is important as we discuss AUC. In previous research (Allison and Johnson, 2011; Nese et al., 2011; Van Hook, 2008), low base rate has been problematic. As Petscher, Kim and Foorman (2011, p. 7) point out, "Whereas sensitivity and specificity are properties of the test itself (Streiner, 2003), sample-based indices are dependent on the proportion of students in the sample that are at risk (i.e., base rate)."

Correlations among Variables

The correlation analysis revealed that all the reading screening measures were significantly correlated with one another and with the outcome measure. As well, all of the demographic variables except free and reduced lunch (FRL), were significantly correlated with the outcome measure (MSP 2011). As expected, FRL and SPED status both demonstrated significant negative correlations with all reading screening measures and prior year MSP (2010). ELL status also had significant negative correlations with prior year MSP and all reading

screening measures with the exception of ORF. Although the percentage of ELL students is low, this lack of correlation with ORF is of interest because similar concerns have been raised in prior research (Valencia et al., 2010). Gender also presented a significant positive correlation for females with the 2010 and 2011 MSP, Gates-MacGinitie (Vocab + Comp) and ORF assessments but not with CBM vocabulary or the TOSRE (Vocab + Comp). This is interesting, as Nese & colleagues (2011) found that gender, ethnicity, special education status and FRL, with the exception of 5th grade FRL, were not significantly associated with the high-stakes assessment in Oregon.

Correlations between ORF and MSP 2010/2011 are of particular interest, as ORF has served individually as the primary screening tool in the majority of previous research (Jenkins et al., 2007). In this study, concurrent validity for ORF with MSP 2010 was $r = .65$, while predictive validity for ORF with MSP 2011 demonstrated a correlation of .45. This means that where 42% of the variance is related between MSP 2010 and ORF, only 20% of the variance is related between MSP 2011 and ORF. While correlation alone is not considered the optimal way to evaluate the predictive capability of a universal screening tool, it does provide initial insight. In and of itself, the correlation between ORF and MSP (concurrent and predictive) while statistically significant, does not seem informative enough for a screening measure.

Additionally, because one purpose of this study was to identify combinations of screening tools that would more accurately predict performance on the MSP, it is also valuable to have information regarding relationships of universal screening tools with one another. Each of the reading measure predictors significantly correlate with MSP 2010 (.65 to .81) and with MSP 2011 (.45 to .64). In addition, the reading measure predictors significantly correlate with one another. Of the reading measure predictors, CBM vocabulary has the lowest correlation with

ORF (.57) and TOSRE (Vocab + Comp) (.63). This translates to 32% of the variance between ORF and 40% of the variance between TOSRE (Vocab + Comp) is shared with CBM vocabulary. This may suggest that CBM vocabulary partially measures a different construct of reading. With this insight, it seems reasonable to pair CBM vocabulary with either ORF or the TOSRE (Vocab + Comp) to increase classification accuracy rate.

Screening Diagnostic Accuracy (AUC)

Another method of evaluating universal screening tools is with receiver operating characteristic (ROC) curve analyses. While this method can provide sensitivity, specificity, positive predictive power and negative predictive power, I chose to report area under the curve (AUC) for each reading measure assessment. As Johnson et al. (2009) describe, "An overall indication of the diagnostic accuracy of a ROC curve is the area under the curve (AUC)" (p.175). A valid and reliable AUC estimate can be interpreted as the probability that the assessment will assign a lower score to a randomly chosen negative example than to a randomly chosen positive example. An AUC greater that .90 is considered excellent; .80 to.90, good; .70 to .80, fair; and below .70, poor.

In this study, proximal measures of reading comprehension such as prior year MSP (.89) and Gates-MacGinitie (Vocab + Comp) (.89), are very good predictors of the outcome measure (MSP 2011). CBM vocabulary (.81) is also a good predictor. Distal measures of reading comprehension such as ORF (.76) and the TOSRE (Vocab + Comp) (.79) are considered "fair" predictors. This pattern of results replicates Johnson et al. (2010) where ORF had an AUC value of .80 in both second and third grade; while the proximal measure of reading comprehension, SAT, had a slightly higher AUC value of .86. It is interesting that the differences in our study are slightly more pronounced than in the study conducted by Johnson and colleagues. This may be

due to the age of the students in each of the study as the students in this study were considerably older than those in the study conducted by Johnson and colleagues.

As suggested by Jenkins & Juel (1993), the relationship of ORF with reading comprehension weakens as students' progress in school. In a study with over 5,000 students in third, fifth, seventh and eighth grade, Silberglitt, Burns, Madyun, and Lail (2006) also found that as students grew older, the relationship between ORF and comprehension weakened. In their study, ORF accounted for 50% of the variance at third grade and only 26% of the variance at eighth grade. Similarly, Schatschneider et al. (2004) found that at third grade text fluency accounted for substantially more variance on the Florida Comprehensive Assessment Test (FCAT) than did reasoning and verbal knowledge; in comparison, at seventh grade, the variance across these assessments were comparable and at tenth grade, reasoning and verbal knowledge dominated ORF.

Predicting Reading Achievement with Prior Year Reading Achievement

Multiple logistic regressions were analyzed in the current study to explore the unique contributions of student demographic characteristics, prior year MSP (2010) and individual reading measures to predict performance on the outcome measure (MSP 2011). In addition, a combination of all predictors was evaluated in an effort to obtain the highest levels of classification accuracy. Findings from this analysis further support the strength of proximal measures of reading comprehension and continues to raise concerns regarding the use of ORF with students in sixth grade.

One of the major findings of this study is that block one of the analysis (prior year MSP, and student demographics) explained 58% of the variance on the MSP 2011. In fact, most of the variance was specific to prior year MSP (2010). This is similar to the findings of several other

studies that examined that prior year assessment with late elementary students (Nese et al., 2011), and middle school students (Denton et al., 2011; Allison & Johnson, 2011). The amount of variance accounted for in these studies ranged from 44% to 70%. Fuchs et al. (2010) said, in reference to middle school students, "It makes more sense to rely on teacher nomination or existing assessment data to identify students…" (p. 24). Evidence from this study certainly helps support this suggestion.

When individual reading measures to the base were added one at a time, ORF, CBM Vocab and TOSRE (Vocab + Comp) each explained only an additional insignificant one percent of the variance beyond the base model. Gates-MacGinitie (Vocab + Comp), however, explained an additional three percent of the variance which is statistically significant. Again, we see evidence that supports the idea that, for older students, measures that more closely related to the outcome measure do better in predicting student success.

In block two of this model, all reading measures (ORF, CBM Vocab, TOSRE (Vocab + Comp)) were added to the analysis together in an effort to create the most robust model, theoretically able to explain the greatest amount of the variance. However, this model was no better than the model with baseline + Gates-MacGinitie (Vocab + Comp) alone. Both of the models accounted for 62% of the variance and had a hit rate of .83. This is valuable information, as efficiency is a concern in universal screening. No additional benefit came from including the fluency and CBM vocabulary measures.

Predicting Reading Achievement without Prior Year Reading Achievement

Because prior year MSP, as anticipated, accounted for such a large percentage of the variance in the outcome measure and because prior year achievement scores are not always

available, the logistic regression process was repeated in a second model without this assessment.

Baseline demographic block. The baseline block (gender, sped, ELL and FRL) without prior year MSP, only accounted for 25% of the variance. While sensitivity (.96) was very high in this block, specificity was extremely low (.38). Of the demographic variables, special education status was a significant predictor along with Gender & ELL status. It was very surprising to see that FRL was not a significant predictor of the MSP 2011 passing status. I anticipated free or reduced lunch to play a bigger role in academic achievement as found in prior studies (Nese et al., 2011) and due to its relation to the prior year MSP score.

Individual reading measures were then added to the model one at a time, in an effort to quantitatively isolate the effectiveness of each measure. First, and foremost, all of the reading measures were significant predictors of the MSP 2011 passing status. In addition, all reading measures maintained the .90 sensitivity level recommended by Jenkins & Colleagues (2007) and overall hit rates were between .76 and .83.

Oral reading fluency. Oral Reading Fluency (ORF) is critical to include in the study for three primary reasons. First, ORF is arguably the most common universal screening tool used in RTI frameworks (Jenkins et al., 2007). Many district leaders have generalized the research conducted at lower grade levels to middle school settings without evaluating the effectiveness of these tools with a different population. Second, fluency measures, by design, are quick to administer and easy to score; this makes them prime candidates for screening measures. However, while these measures only occupy a few minutes per student, total time spent for the teacher can be dramatically different due to the number of students assessed. Finally, the embedded nature of the fluency construct speaks to automaticity and mastery of reading, which

is the goal. It is unclear from previous research if this construct holds value for students in middle school settings.

Oral reading fluency is a distal measure of reading comprehension (Fuchs, L.S, Fuchs D., & Maxwell, 1988). That is to say, it does not directly measure reading comprehension. Users of this assessment assume that a student who reads fluently is capable of transferring that knowledge to overall comprehension of the text and in fact, researchers have demonstrated a strong positive relationship between ORF and reading comprehension for students in elementary school (e.g., Fuchs, Fuchs, Hosp & Jenkins, 2001; Good et al. 2001; Marston, 1989). Several studies refer to "growth deceleration" on reading fluency growth rates (Chall, 1983; Fuchs, Fuchs, Hamlett, Walz, & German, 1993). As Silberglitt et al. (2006, p. 528.) explain in regard to deceleration, "This was probably due to focusing on more complex instructional domains rather than on the basic skills assessed with CBM (Fuchs, Fuchs, Hamlett, Walz, & German, 1993) and is consistent with developmental stage theory research on reading fluency growth rates in general." As students get older, Schilling et al. (2007) speculate that ORF is less closely associated with comprehension perhaps because, state standardized reading tests require increased vocabulary demands and comprehension questions require higher level thinking skills such as the ability to make inferences (Cain & Oakhill, 1999; Yovanoff et al. 2005). This shift in comprehension expectations may impede the effectiveness of fluency as a predictor of reading comprehension success for older readers.

As it turned out, in this study ORF was one of the least effective of the four reading measures in predicting the outcome on the MSP. While statistically significant, ORF only accounted for 34% of the variance when prior year MSP was not available. This inaccuracy is highlighted in the measure specificity for ORF (.45). For the data in this study, this level of

specificity resulted in 56 false positives. In other words, when ORF was used independently, 56 or 57 students were identified as in need of intervention, when in fact they passed the outcome measure. This over-identification is important to schools, as it results in unnecessary expenditure of resources.

That being said, ORF has a high sensitivity level (.92). This level of sensitivity exceeds the .90 threshold recommended by Jenkins and colleagues (2007). In fact, this sensitivity level may position oral reading fluency as a potential first screening measure in a gated-screening procedure. While this sounds like a viable option at first blush, for the data in this study, even with a sensitivity level of .92, ORF would only screen out 61 of the 306 students in the first gate of the screening procedure. Perhaps of greater importance, 16 of the 61 students were false negatives. In other words, 16 students who did not pass the 2011 MSP were not identified by ORF and therein would not be included in the second screen. The second screen concept is designed to reduce the number of false positives, and consequently increase specificity, but it does not help with false negatives or sensitivity. The findings of this research, support the suggestion that the first screen in a two-step process, should have a 99% confidence interval ensuring that 99% of all true positives would be identified in the first screen (Compton et al., 2010). In schools who have a low base rate, that is, a high percentage of the students pass the outcome measure, this could be a beneficial practice. However, in schools with a relatively high base rate, as in this study, the first gate is not likely to eliminate enough students from the risk pool to justify the time necessary for this type of testing logistics.

Vocabulary Curriculum-Based Measurement. It is apparent in the Adam's four-part processer that it is difficult to untangle vocabulary and comprehension into separate assessments, as the processors interact with one another to create meaning. None of the vocabulary and

comprehension assessments in this study truly isolate these constructs from one another. They will, however, be discussed separately since they were entered into the model as separate constructs.

While interest in vocabulary as a construct for screening measures has been growing Jenkins et al. (2007) did not report a single study in which vocabulary was used as a predictor variable. Since then, several studies have used the vocabulary construct as screening measures (Adlof et al., 2010; Compton et al., 2006; Johnson et al. 2009; Nese et al., 2011; Riedel, 2007; Schatschneider, et al., 2004; Speece et al., 2010). One common vocabulary measure used in prior research (Johnson et al., 2009) with younger students is the Peabody Picture Vocabulary Test (PPVT). This assessment of receptive vocabulary is individually-administered and requires a relatively large amount of testing time. This requirement makes tests like the PPVT, inefficient and difficult to use.

In the current study, a CBM Vocab test was used as a screening measure. It is important to distinguish this measure of vocabulary skill from previously mentioned assessments in three primary ways. First, the CBM Vocab assessment is group administered. Second, stimuli for this assessment are not pictures but rather sentences with missing vocabulary words. Third, this is truly a CBM measure; as a result, reliability and validity research are not yet available for this assessment. In recent research, other measures of vocabulary knowledge have been utilized with older students (Nese et al., 2011; Speece et al., 2010; Van Hook 2008).

One finding of the current study was that without prior year MSP in the model, CBM Vocab explained 43% of the variance in the MSP 2011 passing status. While the CBM Vocab assessment had sensitivities above .90, specificity was quite low (.56). In terms of false positives, this translated into 44 false positives using just on the CBM Vocab assessment.

TOSRE (Vocab + Comp) silent reading fluency. Similar to ORF, the TOSRE (Vocab + Comp) is a distal measure of reading comprehension. That is to say, it does not ask students to read an entire passage and answer comprehension questions about what they have read. Instead, the TOSRE (Vocab + Comp) asks students to read a sentence and make a judgment regarding the sentence with a yes/no response. While this task requires a certain level of comprehension, it does not have the same requirements as most outcome measures. However, fluency measures such as ORF, Maze and TOSRE (Vocab + Comp) have been validated as proxies for reading comprehension (Buck & Torgeson, 2003; Ridel, 2007; Roehrig et al., 2008; Shapiro, Solari, & Petscher, 2008).

Results from this study measured ORF, TOSRE (Vocab + Comp), CBM Vocab and the Vocab + Comp in relation to MSP. While the TOSRE (Vocab + Comp) correlated slightly higher than ORF with MSP (.47 and .45 respectively), CBM Vocab and the Gates-MacGinitie (Vocab + Comp) demonstrated an even stronger relationship with the outcome measure (.52 and .64 respectively). When prior year MSP was removed from the base model, TOSRE (Vocab + Comp) had a sensitivity of .91 and specificity of .44 with an overall hit rate of .76. It seems important to note that while TOSRE (Vocab + Comp) explained slightly more of the variance than ORF in this study with 35% and 34% respectively, CBM Vocab and Gates-MacGinitie (Vocab + Comp) did a far better job with 43% and 55% respectively.

Denton et al. (2011) conducted a study with sixth, seventh and eighth grade students using several measures of reading comprehension, oral reading fluency, silent reading fluency, and vocabulary. They found that TOSRE (Vocab + Comp) correlated equally well with the state high-stakes assessment as ORF and GRADE Passage Comprehension. In fact, TOSRE (Vocab + Comp) correlated with WJIII Passage Comprehension better than ORF. In addition, overall

classification accuracy for TOSRE (Vocab + Comp) was dependent on the percentile used for benchmark. The highest classification accuracy achieved for TOSRE (Vocab + Comp) occurred at the 50th percentile and resulted in .69. This classification accuracy rate is slightly lower than the one from the current study.

Gates-MacGinitie (Vocab + Comp). In the current study, when prior year MSP was not part of the model, Gates-MacGinitie (Vocab + Comp) explained the largest portion of the variance, .55 of the reading measures in this study and demonstrated a AUC of .885 on the ROC curve analysis which is considered "good". In addition, Gates-MacGinitie (Vocab + Comp) had the highest overall hit rate .83 on the logistic regression analysis with a sensitivity of .91 and specificity .67. All things considered, Gates-MacGinitie (Vocab + Comp) seems to be a relatively predictive screening tool for students in the sixth grade.

The only other study I could locate that used the Gates-MacGinitie (Vocab + Comp) as a predictor variable was conducted by Speece et al. (2010). Results of this study indicate that no single screening measure adequately identified children as "at-risk". However, a composite of Gates-MacGinitie (Vocab + Comp), Test of Silent Word Reading Fluency and teacher rating was found to explain 46% of the variance and had an AUC of .90. Comparison of these results are of interest, as Speece and colleagues (2010, p. 12) point out, "The results of this study confirm that screening for reading problems in middle childhood requires a multivariate perspective." Results from the current study explained more of the variance using single predictors than did Speece and colleagues composite score. Furthermore, as will be discussed, the addition of screeners does little to improve predictability in the current study.

All predictor variables. In the current study, each of the reading assessments (ORF, CBM Vocab, TOSRE (Vocab + Comp) and Gates-MacGinitie (Vocab + Comp)) were tested for

their predictive utility with and without the prior year MSP included. In the case where prior year MSP is not part of the model, the combined reading measures performed no better than the model with Gates-MacGinitie (Vocab + Comp) alone. Each of the models resulted in an overall hit rate of .83. These findings were surprising because creating composite scores for younger students has had positive results (e.g. Compton et al., 2010: Speece et al., 2010). While this model did not improve classification accuracy, we decided to evaluate two additional, exploratory models: the first model was designed for maximum testing efficiency and the second was an attempt to improve consequential validity.

Model designed for maximum efficiency. One argument for the inclusion of fluency measures as screening tools is their efficiency over longer more comprehensive assessments. In an attempt to bolster the results of ORF, we added the TOSRE (Vocab + Comp) to this model. In doing so, the variance explained by ORF alone was increased from 34-37% when combined with TOSRE (Vocab + Comp). Perhaps of greater importance, however, is that the overall hit rate decreased from .77 to .76 in the combined model in the logistic regression analysis. This seems to suggest that the addition of TOSRE (Vocab + Comp) to ORF is unnecessary.

Another popular measure of silent reading fluency not used in our study, Maze, has been evaluated in several studies. Kim, Petscher, and Foorman (2015, p. 146) stated in regard to silent reading fluency (Maze): “Although silent reading fluency offered somewhat stronger unique explanatory power for students in grades 3-5 than for those in grades 6-10, it consistently added explanatory power.” Decker et al. (2014) also used Maze as a measure of silent reading fluency for students in seventh and eighth grade. They found AUC values of .80 and .77 for seventh and eighth grade respectively. When they combined ORF with Maze they found AUC was increase

to .83 (seventh) and .87 (eighth). These studies offer support for the inclusion of silent reading assessments in screening models.

Model designed for improved consequential validity. Finally, we explored the idea that ORF's consequential validity might be improved through the addition of a measure of vocabulary. To this end, we combined ORF with CBM Vocab. In this case, with prior year MSP excluded from the base model, ORF explained 34% of the variance. The addition of the CBM Vocab measure, this was increased to 45% of the variance. In addition, ORF's overall hit rate of .77 was increased to .80. These increases support the idea that prediction can be improved by adding a measure of vocabulary to ORF.

Schatshneider et al. (2004) research supports the results found in the current study. They found that, students who scored at the lowest level of the FCAT, percentiles for verbal knowledge and reasoning decreased as they got older. Specifically, students in level 1 of the FCAT in 3rd grade scored at the 42nd percentile for verbal reasoning. This percentile decreased to the 30th percentile in 10th grade. This demonstrates that students who struggle with reading comprehension have increased difficulty with verbal knowledge and reasoning as they get older. Therefore, this data supports the idea that measures that tap into the verbal knowledge and reasoning construct may assist in the identification of older students reading comprehension deficits.

Implications for Practice:

Revised Screening Recommendations for Older Students

Based on our findings, educators of sixth grade students may be able to most accurately identify students at-risk of failure on high-stakes assessments by evaluating their performance on the prior year's high-stake assessment score. Use of extant data, such as prior year assessment,

not only saves significant time in testing and resources, but also allows schools to place students in intervention groups at the beginning of the school year. One hurdle in middle school settings, for universal screening, is the difference in constructing student schedules. While elementary students are generally in a single classroom, middle school students often move from subject area to subject area in different classrooms. When universal screening takes place at the beginning of the school year, it can cause a significant disruption as student schedules are modified to accommodate for intervention classes. If prior year test scores are not available, results from this study indicate that the Gates-MacGinitie (Vocab + Comp) is the next best option. While this assessment takes longer to administer than other traditional screening tools, it can be group administered, thus saving evaluator time.

Directions for Future Research

While findings from this study support previous research in regard to the weakening relationship between ORF and reading comprehension as students get older (Jenkins & Jewel 1993; Kim et al., 2015; Silberglitt et al., 2006), it is important to replicate this research with a variety of outcome measures and school demographics. Because outcome measures in prior research have primarily been state high-stakes assessments, it is difficult to generalize the results nationally.

Another consideration for future research should be to examine how base rate of the outcome measure shifts priorities within individual school frameworks. Base rate plays an important role in the feasibility of screening protocols in school districts. That is to say, a school with a low base rate may be able to use a two-gate screening procedure and eliminate a large portion of the total school population from additional assessments in the first gate, thus saving valuable time for diagnoses for the few children who are actually "at-risk." In contrast, schools

with relatively high base rates, as in this study, may not benefit as much from the gated screening procedure as the number of false positives and false negatives may render the first gate insufficient.

Finally, as we explore different constructs of reading comprehension as a means to identify middle school children at risk, it becomes increasingly important to include various tools and combinations of tools designed to capture individual student strengths and needs. Research that includes measures of spelling (Chua, Liow & Yeong, 2014; Foorman & Ciancio, 2005; Maughan, 2009; Speece et al., 2010), vocabulary (Compton et al., 2006; Nese et al., 2011; Schatschneider et al., 2004) and comprehension (Baker et al., 2015; Kim et al., 2015) offer promise for future research.

References

Adams, M. J. (1990). Beginning to read: Learning and thinking about print. *Beginning to read: Learning and thinking about print*.

Adlof, S. M., Catts, H. W., & Lee, J. (2010). Kindergarten predictors of second versus eighth grade reading comprehension impairments. *Journal of Learning Disabilities*, *43*(4), 332-345.

Allison, J. R., & Johnson, E. S. (2011). Identifying Struggling Readers in Middle School with ORF, Maze and Prior Year Assessment Data. *Journal of Educational & Developmental Psychology*, *1*(1).

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999).Standards for educational and psychological testing.Washington, DC: American Educational Research Association.

Anderson, R. C. (1985). Becoming a Nation of Readers: The Report of the Commission on Reading.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. InJ. T. Guthrie (Ed.), Comprehension and teaching: Research reviews (pp. 77-117). Newark. *DE: International Reading Association*.

Anderson, R. C., & Nagy, W. E. (1991). *Word meanings*. Lawrence Erlbaum Associates, Inc.

Aud, S., Hussar, W., Kena, G., Bianco, K., Frohlich, L., Kemp, J., & Tahan, K. (2011). The Condition of Education 2011. NCES 2011-033. *National Center for Education Statistics*.

Badian, N. A. (1999). Persistent arithmetic, reading, or arithmetic and reading disability. *Annals of Dyslexia*, *49*(1), 43-70.

Baker, S. K., Simmons, D. C., & Kame'enui, E. J. (1998). Vocabulary acquisition: Research bases. *What reading research tells us about children with diverse learning needs: Bases and basics*, 183-217.

Beck, I. L., McKeown, M. G., & McCaslin, E. S. (1983). Vocabulary development: All contexts are not created equal. *The Elementary School Journal*, 177-181.

Beck, I. L., Perfetti, C. A., & McKeown, M. G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology*, *74*(4), 506.

Becker, W. C. (1977). Teaching reading and language to the disadvantaged—What we have learned from field research. *Harvard Educational Review*, *47*(4), 518-543.

Bennett, K. J., Lipman, E. L., Brown, S., Racine, Y., Boyle, M. H., & Offord, D. R. (1999). Predicting conduct problems: can high-risk children be identified in kindergarten and grade 1?. *Journal of consulting and clinical psychology*, *67*(4), 470.

Biancarosa, G., & Snow, C. E. (2004). *Reading next: A vision for action and research in middle and high school literacy: A report from Carnegie Corporation of New York*. Alliance for Excellent Education.

Bowers, P. N., & Kirby, J. R. (2010). Effects of morphological instruction on vocabulary acquisition. *Reading and Writing*, *23*(5), 515-537.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual review of psychology*, *53*(1), 371-399.

Braze, D., Tabor, W., Shankweiler, D. P., & Mencl, W. E. (2007). Speaking up for vocabulary reading skill differences in young adults. *Journal of learning disabilities*, *40*(3), 226-243.

Buck, J., Torgesen, J., & Schatschneider, C. *Predicting FCAT-SSS scores using prior performance on the FCAT-SSS, FCAT-NRT, and SAT9* (Vol. 4). FCRR Technical Report.

Catts, H. W., Compton, D., Tomblin, J. B., & Bridges, M. S. (2012). Prevalence and nature of late-emerging poor readers. *Journal of educational psychology*, *104*(1), 166.

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the Risk of Future Reading Difficulties in Kindergarten Children A Research-Based Model and Its Clinical Implementation. *Language, speech, and hearing services in schools*, *32*(1), 38-50.

Catts, H. W., & Hogan, T. P. (2002). The fourth grade slump: Late emerging poor readers. In *Poster presented at the annual conference of the Society for the Scientific Study of Reading, Chicago, IL*.

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2008). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities*.

Chall, J. S., & Jacobs, V. A. (2003). Poor children's fourth-grade slump. *American educator*, *27*(1), 14-17.

Christ, T. J., & Silberglitt, B. (2007). Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency. *School Psychology Review*.

Chua, S. M., Liow, S. J. R., & Yeong, S. H. (2014). Using Spelling to Screen Bilingual Kindergarteners At Risk for Reading Difficulties. *Journal of learning disabilities*, 0022219414538519.

Compton, D. L. (2006). How should “unresponsiveness” to secondary intervention be operationalized? It is all about the nudge. *Journal of learning disabilities*, *39*(2), 170-173.

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., ... & Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology*, *102*(2), 327.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*, *98*(2), 394.

Compton, D. L., Fuchs, D., Fuchs, L. S., Elleman, A. M., & Gilbert, J. K. (2008). Tracking children who fly below the radar: Latent transition modeling of students with late-emerging reading disability. *Learning and Individual Differences*, *18*(3), 329-337.

Compton, D. L., Gilbert, J. K., Jenkins, J. R., Fuchs, D., Fuchs, L. S., Cho, E., ... & Bouton, B. (2012). Accelerating Chronically Unresponsive Children to Tier 3 Instruction What Level of Data Is Necessary to Ensure Selection Accuracy?. *Journal of learning disabilities*, *45*(3), 204-216.

Compton, D. L., Olson, R. K., DeFries, J. C., & Pennington, B. F. (2002). Comparing the relationships among two different versions of alphanumeric rapid automatized naming and word level reading skills. *Scientific Studies of Reading*, *6*(4), 343-368

Coyne, M.D., Kame'enui, E.J., & Carnine, D.W. (2007). *Effective teaching strategies that accommodate diverse learners* (3rd ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297-334.

Cunningham, A. E., & Stanovich, K. E. (1998). The impact of print exposure on word recognition.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, 37-54

Daniel, S. S., Walsh, A. K., Goldston, D. B., Arnold, E. M., Reboussin, B. A., & Wood, F. B. (2006). Suicidality, school dropout, and reading problems among adolescents. *Journal of Learning Disabilities*, *39*(6), 507-514.

Davis, F. B. (1944). The interpretation of frequency ratings obtained from" The Teachers Word Book.". *Journal of Educational Psychology*, *35*(3), 169.

Davis, F. B. (1944). Fundamental factors of comprehension in reading. *Psychometrika*, *9*(3), 185-197.

Davis, F. B. (1968). Research in comprehension in reading. *Reading Research Quarterly*, 499-545.

Davis, G. N., Lindo, E. J., & Compton, D. L. (2007). Children at Risk for Reading Failure; Constructing an Early Screening Measure. *Teaching Exceptional Children*, *39*(5), 32-37.

Dearing, E., Berry, D., & Zaslow, M. (2006). Poverty during early childhood. *Blackwell handbook of early childhood development*, 399-423.

Dennis, D. V. (2013). Heterogeneity or Homogeneity What Assessment Data Reveal About Struggling Adolescent Readers. *Journal of Literacy Research*, *45*(1), 3-21.

Deno, S. L. (1989). Curriculum-based measurement and special education services: A fundamental and direct relationship.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education*, *37*(3), 184-192.

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional children*.

Denton, C. A., Barth, A. E., Fletcher, J. M., Wexler, J., Vaughn, S., Cirino, P. T., ... & Francis, D. J. (2011). The relations among oral and silent reading fluency and comprehension in middle school: Implications for identification and instruction of students with reading difficulties. *Scientific Studies of Reading*, *15*(2), 109-135.

Denton, C. A., Fletcher, J. M., Anthony, J. L., & Francis, D. J. (2006). An evaluation of intensive intervention for students with persistent reading difficulties. *Journal of Learning Disabilities*, *39*(5), 447-466.

Duncan, G. J., & Brooks-Gunn, J. (2000). Family poverty, welfare reform, and child development. *Child development*, *71*(1), 188-196.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd Edition) Bloomington, MN: Pearson Assessments.

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, A., Tackett, K. K., & Schnakenberg, J. W. (2009). A synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers. *Review of Educational Research*, *79*(1), 262-300.

Elliott, S. N., Huai, N., & Roach, A. T. (2007). Universal and early screening for educational difficulties: Current and future approaches. *Journal of School Psychology*, *45*(2), 137-161.

Evangelista, L. S., Rasmusson, K. D., Laramee, A. S., Barr, J., Ammon, S. E., Dunbar, S., ... & Yancy, C. W. (2010). Health literacy and the patient with heart failure—implications for patient care and research: a consensus statement of the Heart Failure Society of America. *Journal of cardiac failure*, *16*(1), 9-16.

Flanagan, K. S., Bierman, K. L., & Kam, C. M. (2003). Identifying at-risk children at school entry: The usefulness of multibehavioral problem profiles. *Journal of Clinical Child and Adolescent Psychology*, *32*(3), 396-407.

Fletcher, J. M., Francis, D. J., Shaywitz, S. E., Lyon, G. R., Foorman, B. R., Stuebing, K. K., & Shaywitz, B. A. (1998). Intelligent testing and the discrepancy model for children with learning disabilities. *Learning Disabilities Research & Practice*.

Foorman, B. R., & Ciancio, D. J. (2005). Screening for Secondary Intervention Concept and Context. *Journal of Learning Disabilities*, *38*(6), 494-499.

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology*, *90*(1), 37.

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2010). Rethinking response to intervention at middle and high school. *School Psychology Review*, *39*(1), 22-28.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Allinder, R. M. (1991). Effects of expert system advice within curriculum-based measurement on teacher planning and student achievement in spelling. *School Psychology Review*.

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

Fuchs, L.S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education*, *9*(2), 20–28.

Garbarino, J., & Ganzel, B. (2000). The human ecology of early risk. *Handbook of early childhood intervention*, *2*, 76-93.

Gates-MacGinitie Reading Tests (1978). Boston: Houghton Mifflin.

Gerald, D. E. (1999). *Projections of education statistics to 2008*. DIANE Publishing.

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). Assisting Students Struggling with Mathematics: Response to Intervention (RtI) for Elementary and Middle Schools. NCEE 2009-4060. *What Works Clearinghouse*.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology*, *45*(2), 117-135.

Good, R. H., Kaminski, R. A., Smith, S., Laimon, D., & Dill, S. (2004). *Dynamic indicators of basic early literacy skills*. Sopris West Educational Services.

Good III, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, *5*(3), 257-288.

Gordon-Larsen, P., Harris, K. M., Ward, D. S., & Popkin, B. M. (2003). Acculturation and overweight-related behaviors among Hispanic immigrants to the US: the National Longitudinal Study of Adolescent Health. *Social science & medicine*, *57*(11), 2023-2034.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics.* New York: Wiley.

Haager, D. E., Klingner, J. E., & Vaughn, S. E. (2007). *Evidence-based reading practices for response to intervention*. Paul H Brookes Publishing.

Hasbrouck, J., & Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher*, *59*(7), 636-644.

Hattie, J. (2013). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hosp, J. L., & Ardoin, S. P. (2008). Assessment for instructional planning. *Assessment for Effective Intervention*, *33*(2), 69-77.

Howe, K. B., & Shinn, M. M. (2002). Standard reading assessment passages (RAPs) for use in general outcome measurement: A manual describing development and technical features. *Eden Prairie, MN: Edformation*.

Individuals with Disabilities Education Act, 20 U.S.C. § 1401 *et seq.* (1997).

Individuals with Disabilities Education Improvement Act, H.R. 1350, 108th Congress (2004).

Individuals with Disabilities Education Act. (2010). 20 U.S.C §§1401 et seq.

Jenkins, J. R. (2003, December). *Candidate measures for screening at-risk students*. Paper presented at the National Research Center on Learning Disabilities Responsiveness-to-Intervention symposium, Kansas City, MO. Retrieved April 3, 2006, from http://www.nrcld.org/symposium2003/jenkins/index.html.

Jenkins, J. R., Fuchs, L. S., Van Den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, *95*(4), 719.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review*, *36*(4), 582.

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

Jenkins, J. R., & O'Connor, R. E. (2002). Early identification and intervention for young children with reading/learning disabilities. *Identification of learning disabilities: Research to practice*, 99-149.

Jenkins, J. R., Schiller, E., Blackorby, J., Thayer, S.K., & Tilly, W.D. (2013). Responsiveness to intervention in reading: Architecture and practices. *Learning Disability Quarterly, 36,* 36-46. Doi:10.1177/0731948712464963

Johnson, E. S., Humphrey, M., Mellard, D. F., Woods, K., & Swanson, H. L. (2010). Cognitive processing deficits and students with specific learning disabilities: A selective meta-analysis of the literature. *Learning Disability Quarterly*, *33*(1), 3-18.

Johnson, E. S., Jenkins, J. R., & Petscher, Y. (2010). Improving the accuracy of a direct route screening process. *Assessment for Effective Intervention*, *35*(3), 131-140.

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice*, *24*(4), 174-185.

Johnson, E., Mellard, D. F., Fuchs, D., & McKnight, M. A. (2006). Responsiveness to Intervention (RTI): How to Do It.[RTI Manual]. *National Research Center on Learning Disabilities*.

Johnston, P. H. (2011). Response to intervention in literacy: Problems and possibilities. *The Elementary School Journal*, *111*(4), 511-534.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of educational Psychology*, *80*(4), 437.

Juel, C. (1996). What makes literacy tutoring effective?. *Reading Research Quarterly*, *31*(3), 268-289.

Kamhi, A. G., & Catts, H. W. (2012). *Language and reading disabilities*. Pearson.

Kieffer, M. J. (2010). Socioeconomic status, English proficiency, and late-emerging reading difficulties. *Educational Researcher*, *39*(6), 484-486.

Kogan, M. D., Blumberg, S. J., Schieve, L. A., Boyle, C. A., Perrin, J. M., Ghandour, R. M., ... & van Dyck, P. C. (2009). Prevalence of parent-reported diagnosis of autism spectrum disorder among children in the US, 2007. *Pediatrics*, *124*(5), 1395-1403.

Kutner, M., Greenberg, E., Jin, Y., Boyle, B., Hsu, Y. C., & Dunleavy, E. (2007). Literacy in Everyday Life: Results from the 2003 National Assessment of Adult Literacy. NCES 2007-490. *National Center for Education Statistics*.

Kutner, M., Greenburg, E., Jin, Y., & Paulsen, C. (2006). The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy. NCES 2006-483. *National Center for Education Statistics*.

Lang, L., Torgesen, J., Vogel, W., Chanter, C., Lefsky, E., & Petscher, Y. (2009). Exploring the relative effectiveness of reading interventions for high school students. *Journal of Research on Educational Effectiveness*, *2*(2), 149-175.

Leach, J. M., Scarborough, H. S., & Rescorla, L. (2003). Late-emerging reading disabilities. *Journal of Educational Psychology*, *95*(2), 211.

Lee, B. A., Tyler, K. A., & Wright, J. D. (2010). The new homelessness revisited. *Annual Review of Sociology*, *36*, 501.

Lichtenstein R, Ireton H. (1984) *Pre-school Screening: Identifying Young Children with Developmental and Educational Problems.* New York: Grune Stratton.

Lipka, O., Lesaux, N. K., & Siegel, L. S. (2006). Retrospective analyses of the reading development of grade 4 students with reading disabilities risk status and profiles over 5 years. *Journal of Learning Disabilities*, *39*(4), 364-378.

Marston, D. B. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. *Curriculum-based measurement: Assessing special children*, *1*, 18-78.

Maughan, B., Messer, J., Collishaw, S., Pickles, A., Snowling, M., Yule, W., & Rutter, M. (2009). Persistence of literacy problems: spelling in adolescence and at mid-life. *Journal of Child Psychology and Psychiatry*, *50*(8), 893-901.

McGill-Franzen, A. (1987). Failure to learn to read: Formulating a policy problem. *Reading Research Quarterly*, 475-490.

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

McKeown, M. G., Beck, I. L., Omanson, R. C., & Perfetti, C. A. (1983). The effects of long-term vocabulary instruction on reading comprehension: A replication. *Journal of Literacy Research*, *15*(1), 3-18.

McLoyd, V. C. (1998). Socioeconomic disadvantage and child development. *American psychologist*, *53*(2), 185.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, *18*(2), 5-11.

Metz, C. E. (1978, October). Basic principles of ROC analysis. In *Seminars in nuclear medicine* (Vol. 8, No. 4, pp. 283-298). WB Saunders.

Meyer, B.J.F, McConkie, G.W. (1973) What is recalled after hearing a passage? *Journal of Educational Psychology*. 65,109–117

Morsy, L., Kieffer, M., & Snow, C. (2010). Measure for Measure: A Critical Consumers' Guide to Reading Comprehension Assessments for Adolescents. Final Report from Carnegie Corporation of New York's Council on Advancing Adolescent Literacy. *Carnegie Corporation of New York*.

National Association of State Directors of Special Education & Council of Administrators of Special Education (2006, May) *Response to intervention: NASDSE and CASE white paper on RTI.* Alexandria, VA: Author.

National Center on Response to Intervention. (2010, March). *Essential components of RTI-closer look at response to intervention.* Washington, DC: U.S. Department of Education, Office of Special Education Programs, Author. Retrieved from http://www.cldinternational.org/Articles/rtiessentialcomponents.pdf

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

O'Connor, R. E., Bocian, K., Beebe-Frankenberger, M., & Linklater, D. L. (2010). Responsiveness of students with language difficulties to early intervention in reading. *The Journal of Special Education*, *43*(4), 220-235.

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

O'Connor, R. E., & Klingner, J. (2010). Poor responders in RTI. *Theory Into Practice*, *49*(4), 297-304.

Ofiesh, N., & Mather, N. (2013). Resilience and the child with learning disabilities. In *Handbook of resilience in children* (pp. 329-348). Springer US.

Parrila, R., Aunola, K., Leskinen, E., Nurmi, J. E., & Kirby, J. R. (2005). Development of Individual Differences in Reading: Results From Longitudinal Studies in English and Finnish. *Journal of Educational Psychology*, *97*(3), 299.

Pearson Education. (2012). AIMSweb. San Antonio, TX. Retrieved from http://www.aimsweb.com

Pencina, M. J., D'Agostino, R. B., & Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine*, *27*(2), 157-172.

Pepe, M. S., Janes, H., Longton, G., Leisenring, W., & Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American journal of epidemiology*, *159*(9), 882-890.

Petscher, Y., Kim, Y. S., & Foorman, B. R. (2011). The importance of predictive power in early screening assessments: Implications for placement in the response to intervention framework. *Assessment for Effective Intervention*, 1534508410396698.

Pike, A., Iervolino, A. C., Eley, T. C., Price, T. S., & Plomin, R. (2006). Environmental risk and young children's cognitive and behavioral development. *International Journal of Behavioral Development*, *30*(1), 55-66.

Pressley, M., Hilden, K., & Shankland, R. (2005). *An evaluation of end-Grade-3 Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Speed reading without comprehension, predicting little* (technical report).

Pyle, N., & Vaughn, S. (2012). Remediating reading difficulties in a response to intervention model with secondary students. *Psychology in the Schools*, *49*(3), 273-284.

Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly*, *42*(4), 546-567.

Ritchey, K. D., Silverman, R. D., Montanaro, E. A., Speece, D. L., & Schatschneider, C. (2012). Effects of a tier 2 supplemental reading intervention for at-risk fourth-grade students. *Exceptional children*, *78*(3), 318-334.

Ritchey, K. D., & Speece, D. L. (2004). Early identification of reading disabilities: Current status and new directions. *Assessment for Effective Intervention*, *29*(4), 13-24.

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology*, *46*(3), 343-366.

Satz, P., & Fletcher, J. M. (1981). Emergent trends in neuropsychology: An overview. *Journal of Consulting and Clinical Psychology*, *49*(6), 851.

Scammacca, N., Roberts, G., Vaughn, S., Edmonds, M., Wexler, J., Reutebuch, C. K., & Torgesen, J. K. (2007). Interventions for Adolescent Struggling Readers: A Meta-Analysis with Implications for Practice. *Center on Instruction*.

Scammacca, N. K., Roberts, G., Vaughn, S., & Stuebing, K. K. (2013). A meta-analysis of interventions for struggling readers in Grades 4–12: 1980–2011. *Journal of learning disabilities*, 0022219413504995.

Scarborough, H. S. (1998). Predicting the future achievement of second graders with reading disabilities: Contributions of phonemic awareness, verbal memory, rapid naming, and IQ. *Annals of Dyslexia*, *48*(1), 115-136.

Schatschneider, C. (2006). *Reading difficulties: Classification and issues of prediction.* Paper presented at the Pacific Coast Regional Conference, San Diego, CA.

Schatschneider, C., Buck, J., Torgesen, J., Wagner, R., Hassler, L., Hecht, S., & Powell-Smith, K. (2004). A Multivariate Study of Individual Differences in Performance on the Reading Portion of the Florida Comprehensive Assessment Test: A Brief Report. *Florida Center for Reading Research*.

Schilling, S. G., Carlisle, J. F., Scott, S. E., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *The Elementary School Journal*, *107*(5), 429-448.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, *96*(4), 523.

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review*, *78*(1), 40-59.

Shapiro, E. S., Solari, E., & Petscher, Y. (2008). Use of a measure of reading comprehension to enhance prediction on the state high stakes assessment. *Learning and Individual Differences*, *18*(3), 316-328.

Shinn, M. M., and M. R. Shinn. *"AIMSweb training workbook: Administration and scoring of reading curriculum-based measurement (R-CBM) for use in general outcome measurement." Eden Prairie, MN: Edformation, Inc. Available at: www. aimsweb. com/uploads/files/adminandscoringrcbm09292005. pdf* (2002).

Shinn, M. R. (Ed.). (1989). *Curriculum-based measurement: Assessing special children*. Guilford Press.

Shippen, M. E., Houchins, D. E., Crites, S. A., Derzis, N. C., & Patterson, D. (2010). An examination of the basic reading skills of incarcerated males. *Adult Learning*, *21*(3/4), 4-12.

Singer, H. (1965). A developmental model for speed of reading in grades three through six. *Reading Research Quarterly*, 29-49.

Simmons, D. C., Coyne, M. D., Kwok, O. M., McDonagh, S., Harn, B. A., & Kame'enui, E. J. (2008). Indexing response to intervention a longitudinal study of reading risk from kindergarten through third grade. *Journal of Learning Disabilities*, *41*(2), 158-173.

Snyder, H. N., & Sickmund, M. (1995). *Juvenile offenders and victims: a national report: preview*. DIANE Publishing.

Snyder, H. N., & Sickmund, M. (2006). Juvenile offenders and victims: 2006 national report. *Office of juvenile justice and delinquency prevention*.

Speece, D. L. (2005). Hitting the Moving Target Known as Reading Development Some Thoughts on Screening Children for Secondary Interventions. *Journal of Learning Disabilities*, *38*(6), 487-493.

Speece, D.L. (2012). Curriculum-Based Measurement progress monitoring and health of general education. In S. Rose, C.Espin, K. McMaster, & M. Wayman (Eds.), *A measure of success: The influence of Curriculum-Based Measurement on education* (pp. 179-184). Minneapolis: University of Minnesota Press.

Speece, D. L., Ritchey, K. D., Silverman, R., Schatschneider, C., Walker, C. Y., & Andrusik, K. N. (2010). Identifying children in middle childhood who are at risk for reading problems. *School psychology review*, *39*(2), 258.

Speece, D. L., Schatschneider, C., Silverman, R., Case, L. P., Cooper, D. H., & Jacobs, D. M. (2011). Identification of reading problems in first grade within a response-to-intervention framework. *The Elementary school journal*, *111*(4), 585.

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review*, *30*(3), 407-419.

Stage, S. A., Sheppard, J., Davidson, M. M., & Browning, M. M. (2001). Prediction of first-graders' growth in oral reading fluency using kindergarten letter fluency. *Journal of School Psychology*, *39*(3), 225-237.

Stanovich, K. E. (2013). The Impact of Print Exposure on Word Recognition. *Word Recognition in Beginning Literacy*, 235.

Steadman, H. J., Silver, E., Monahan, J., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., ... & Banks, S. (2000). A classification tree approach to the development of actuarial violence risk assessment tools. *Law and human behavior*, *24*(1), 83-100.

Sternberg, R. J. (1987). The psychology of verbal comprehension. *Advances in instructional psychology*, *3*, 97-151.

Swanson, E., Wanzek, J., Haring, C., Ciullo, S., & McCulley, L. (2013). Intervention fidelity in special and general education research journals. *The Journal of Special Education*, *47*(1), 3-13.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological science in the public interest*, 1-26.

Thurstone, L. L. (1946). Note on a reanalysis of Davis' reading tests. *Psychometrika*, *11*(3), 185-188.

Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research & Practice*, *15*(1), 55-64.

Torgesen J, Nettles S, Howard P, Winterbottom R. (2003) FCRR Report No.6. Tallahassee, FL: Florida Center for Reading Research at Florida State University. Brief report of a study to investigate the relationship between several brief measures of reading fluency and performance on the Florida Comprehensive Assessment Test–Reading in 4th, 6th, 8th, and 10th grades. Retrieved from http://www.fcrr.org/TechnicalReports/Progress_monitoring_report.pdf.

Torgesen, J., Nettles, S., Howard, P., & Winterbottom, R. (2005). Brief Report of a Study to investigate the relationship between several brief measures of reading fluency and performance on the Florida Comprehensive Assessment Test-Reading in 4th, 6th, 8th, and 10th grades. (Technical Report #6). Tallahassee, FL: Florida Center for Reading Research

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1997). Prevention and remediation of severe reading disabilities: Keeping the end in mind. *Scientific studies of reading*, *1*(3), 217-234.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Rose, E., Lindamood, P., Conway, T., & Garvan, C. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology*, *91*(4), 579.

Tsien, C. L., Fraser, H. S., Long, W. J., & Kennedy, R. L. (1998). Using classification tree and logistic regression methods to diagnose myocardial infarction. *Studies in health technology and informatics*, (1), 493-497.

Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly*, *45*(3), 270-291.

VanHook, J. A. (2008) *The Reliability and Validity of Screening Measures in Reading (*Doctoral Dissertation) Retrieved from Louisiana State University Electronic Thesis & Dissertation Collection. Etd-03052008-184657.

Vaughn, S., Cirino, P. T., Wanzek, J., Wexler, J., Fletcher, J. M., Denton, C. D. & Francis, D. J. (2010). Response to intervention for middle school students with reading difficulties: Effects of a primary and secondary intervention. *School Psychology Review*, *39*(1), 3.

Vaughn, S., Denton, C. A., & Fletcher, J. M. (2010). Why intensive interventions are necessary for students with severe reading difficulties. *Psychology in the Schools*, *47*(5), 432-444.

Vaughn, S., Fletcher, J. M., Francis, D. J., Denton, C. A., Wanzek, J., Wexler, J. & Romain, M. A. (2008). Response to intervention with older students with reading difficulties. *Learning and Individual Differences*, *18*(3), 338-345.

Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning disabilities research & practice*, *18*(3), 137-146.

Vaughn, S., Wexler, J., Leroux, A., Roberts, G., Denton, C., Barth, A., & Fletcher, J. (2012). Effects of intensive reading intervention for eighth-grade students with persistently inadequate response to intervention. *Journal of Learning Disabilities*, *45*(6), 515-525.

Wagner, D. A. (2000). EFA 2000 Thematic Study on Literacy and Adult Education.

Wagner, M., Newman, L., Cameto, R., Garza, N., & Levine, P. (2005). After high school: A first look at the postschool experiences of youth with disabilities. A report from the National Longitudinal Transition Study-2 (NLTS2). *Online Submission*.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2010). *Test of silent reading efficiency and comprehension*. Pro-Ed.

Wanzek, J., & Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *School Psychology Review*, *36*(4), 541.

Wanzek, J., Vaughn, S., Roberts, G., & Fletcher, J. M. (2011). Efficacy of a reading intervention for middle school students with learning disabilities. *Exceptional children*, *78*(1), 73-87.

Wanzek, J., Vaughn, S., Scammacca, N. K., Metz, K., Murray, C. S., Roberts, G., & Danielson, L. (2013). Extensive reading interventions for students with reading difficulties after grade 3. *Review of Educational Research*, 0034654313477212.

Wanzek, J., Wexler, J., Vaughn, S., & Ciullo, S. (2010). Reading interventions for struggling readers in the upper elementary grades: A synthesis of 20 years of research. *Reading and writing*, *23*(8), 889-912.

Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education*, *41*(2), 85-120.

Wood, D. E. (2006). Modeling the relationship between oral reading fluency and performance on a statewide reading test. Educational Assessment, 11, 85-104

US Department of Health and Human Services. (2007). health resources and services Administration. *Measuring success for healthy people 2010: National agenda for children with special health care needs*.

Yell, M. L., Shriner, J. G., & Katsiyannis, A. (2006). Individuals with disabilities education improvement act of 2004 and IDEA regulations of 2006: Implications for educators, administrators, and teacher trainers. *Focus on exceptional children*, *39*(1), 1-24.

Zhou, Xiao-Hua, Nancy A. Obuchowski, and Donna K. McClish. "Analysis of correlated ROC data." *Statistical Methods in Diagnostic Medicine* (2002): 274-306.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, *39*(4), 561-577.