2015

Highly Decodable Reading Passages as a First-Grade Screening Measure: A Validation Study

Kirra Guard
Lehigh University

Highly Decodable Reading Passages as a First-Grade Screening Measure: A Validation Study

by

Kirra B. Guard

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Doctor of Philosophy

in

School Psychology

Lehigh University

March 2015

Certificate of Approval

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Date

_____
Edward S. Shapiro, Ph.D.
Dissertation Director
Professor of School Psychology

_____
Accepted Date

Committee Members:

_____
Robin Hojnoski, Ph.D.
Associate Professor of School Psychology

_____
Mark R. Shinn, Ph.D.
Professor of School Psychology and
Program Coordinator,
National Louis University

_____
Mary Beth Calhoon, Ph.D.
Associate Professor of Special Education,
University of Miami

Acknowledgements

First and foremost, I would like to thank my advisor, mentor, and dissertation committee chair, Dr. Edward Shapiro, for the support and guidance he has provided me over the last six years. In the journey to this destination Dr. Shapiro has provided me with challenging and rewarding experiences that have focused my interests in promoting success for all children, while opening my eyes to the potential we have as educators and researchers to change the lives of young people and their families. I am truly thankful for the faith he has had in me, and the motivation he has provided to keep me moving. The greatest reward as one of Dr. Shapiro's mentees is to make him proud, both as a student, and as a graduate. I hope that I can do this, as a tribute to the memories and lessons he has given me.

I would also like to express my sincere gratitude to the members of my dissertation committee, Dr. Robin Hojnoski, Dr. Mark Shinn, and Dr. Mary Beth Calhoon. I am ever grateful for the time, patience, and thoughtfulness you have shown me throughout the dissertation process. In particular, I am humbled by Dr. Shinn's willingness to allow me to conduct this research on his Highly Decodable Passages, and, along the way, to answer my many questions and to spend time offering thought-provoking suggestions and questions for improving my research. But most especially, I am thankful to each one of you, giants in the field of education, for having shepherded me through this process and for helping me to appreciate the innovation, curiosity, and thoughtful practice that education requires.

Finally, I would like to thank my family and friends for the love and support they have shown over the past six years. Lehigh has brought some of the dearest friends I have into my life. It is difficult to express the bond that this program has helped me forge with some of my fellow students and faculty over the years. I can only say that I am truly blessed to have been given the memories and friendships I have, and that I will leave Lehigh with a sadness that things will never quite be the same. At the same time, the road to this milestone has been paved with kindness, forgiveness, and endless patience from friends and family, most especially my husband. Thank you for countless meals, dog-walks, words of encouragement, hugs, and days of tolerating my physical and emotional absence. Most of all, though, thank you for believing in me when I could not always believe in myself. I love you.

1 Title Page……i
2 Copyright Page……ii
3 Approval Page……iii
4 Acknowledgements……iv
5 Table of Contents……v
6 List of Tables……vii
7 List of Figures……viii
8 Abstract……1
9 Chapter I: Statement of the Problem……2
9.1 The Importance of Early Identification and Remediation……3
9.2 Curriculum-Based Measurement in Universal Screening……5
9.3 Reading CBM……6
9.4 Scientific Issues with Current Reading Screening Measures for Early Elementary Students……10
9.4.1 Scientific Issues with Reading CBM……10
9.4.2 Scientific Issues with Other Reading Screening Measures……11
9.5 Practical Issues with Current Reading Screening Measures for Early Elementary Students……15
9.5.1 Efficiency Issues……15
9.5.2 Acceptability Issues……16
9.6 Possible Alternatives to Current Reading Screening Measures for Early Elementary Students……18
9.6.1 Word Identification Fluency……18
9.6.2 Decodable Text……19
9.6.3 Highly Decodable Passages……22
9.7 Research Questions……23
10 Chapter II: Review of the Literature……24
10.1 Reading Processes and Reading Development……24
10.1.1 Theories of Reading……24
10.1.2 Stages of Reading Development……29
10.1.3 Empirical Support for the Importance of Automaticity……32
10.2 Curriculum-Based Measurement……35
10.2.1 Reading Curriculum-Based Measures……36
10.2.2 Early Literacy Curriculum-Based measures……38
10.2.3 Empirical Support for R-CBM……39
10.3 Scientific Issues with Current Curriculum-Based Measures for Screening Early Readers……45
10.3.1 Scientific Issues with Reading CBM……45
10.3.2 Scientific Issues with Other Curriculum-Based Measures……46
10.4 Practical Issues with Current Curriculum-Based Measures for Screening Early Readers……48
10.5 Possible Alternatives to Current Curriculum-Based Measures for Screening Early Readers……51
10.5.1 Word Identification Fluency……51
10.5.2 Decodable Text……53
11 Chapter III: Methods……57
11.1 Participants……57
11.2 Setting……59
11.3 Measures……60
11.4 Procedures……63
11.5 Data Analyses……65

11.5.1 A Priori Analyses…….....65
11.5.2 Preliminary Analyses……66
11.5.3 Analyses of Research Questions……...67
12 Chapter IV: Results……73
12.1 Preliminary Analyses……..……..…...73
12.2 Analysis of Research Questions……..……...……..74
12.2.1 Reliability……..74
12.2.2 Convergent Validity……..75
12.2.3 Hierarchical Linear Modeling……...76
12.2.4 Diagnostic Accuracy……77
12.2.5 Acceptability……80
13 Chapter V: Discussion……82
13.1 Reliability Results……..……..…83
13.2 Convergent Validity Results……..……..…84
13.3 HLM Results……..……..85
13.4 Diagnostic Accuracy Results……...……87
13.5 Teacher Acceptability Results……...……...91
13.6 Limitations……92
13.7 Implications for Practice……...……94
13.8 Future Research Directions……..……96
13.9 Conclusions……..……...99
14 References…….....101
15 Appendix A……113
16 Tables and Figures……116
17 Curriculum Vitae..……122

List of Tables

Table 1……..……..……..**116**
Descriptive Statistics for Student-Administered Measures in the Winter and Spring
Table 2……**117**
Reliability Correlation Matrices Between HD Passages for Test-Retest and Alternate Form Reliability at Each Assessment Period
Table 3……**118**
Convergent Validity Correlation Matrix
Table 4……**119**
Fixed Effects Estimates for HLM
Table 5……**119**
Random Effects Estimates for HLM

List of Figures

Figure 1……..……..……120
Plot of ROC curve predicting to the 25th percentile on the GRADE Comprehension Composite
Figure 2……...121
Plot of ROC curve predicting to the 40th percentile on the GRADE Comprehension Composite

Abstract

Early identification and intervention is essential for promoting achievement in early readers and preventing long-term reading difficulties (Cunningham & Stanovich, 1997; Juel, 1988; Oakhill & Cain, 2012; Spira, Bracken, & Fischel, 2005). Universal screening represents a widely accepted practice for identifying students in need of intervention (Fuchs & Vaughn, 2012). However, existing screening measures demonstrate a number of scientific and practical limitations, such as floor effects, poor predictive accuracy, and limited face validity, and can also be time consuming to administer with multiple measures in kindergarten and first grade (e.g., Catts et al.; 2009; Clemens, Hilt-Panahon, Shapiro, & Yoon, 2012; Goffreda, DiPerna, & Pedersen, 2009; Johnson, Jenkins, Petscher, & Catts, 2009, Goodman, 2006; Pearson, 2006). A newly developed screening measure for early readers, Highly Decodable Passages (HD passages, Shinn, 2009; 2012) was developed in response to these issues.

The current study was intended to investigate the psychometric properties, as well as the acceptability of HD passages. A total of 234 first grade students from 4 elementary schools in Eastern Pennsylvania participated in the study. A group of 20 first grade teachers in Pennsylvania and New York participated in an acceptability survey. Students were assessed in the winter and spring of first grade using HD passages and screening procedures adopted by each school (DIBELS Next; Good et al., 2013). In the spring, students were administered a standardized criterion outcome measure (GRADE; Williams, 2001). Teachers completed an electronic acceptability survey online. Results indicate strong reliability, validity, and diagnostic accuracy, as well as an influence of classroom membership on HD passage outcome scores. Results of the acceptability survey failed to indicate a significant difference between teacher opinions of HD passages versus existing measures of nonsense word fluency.

Chapter I: Statement of the Problem

A plethora of evidence suggests that students in the United States are not developing the skills they need to be proficient readers. According the most recent results of the National Assessment of Educational Progress [NAEP; National Center for Education Statistics (NCES), 2011], only about 34% of students nationwide were able to demonstrate *proficient* or *advanced* performance on assessments of reading. For fourth graders, this result represents a non-significant increase of 1 percentage point from the previous assessments in 2007 and 2009, while eighth-grade students demonstrated a significant increase of 2 percentage points from 2009 assessments. Looking back over the past 9 years, however, indicates that students in the United States have made very little growth in reading achievement. Scores for fourth graders have increased 3 points since 2002, while growth in eighth graders' performance has improved only 1 percentage point since that time.

Overall, these statistics paint a grim picture of literacy development for students in American schools, one that is rather surprising given our nation's general prosperity and image as a global power. Additionally, compared to other countries around the world, US students demonstrate underwhelming reading performance. In an international evaluation of reading performance across 64 countries, 15-year-old US students demonstrated an average proficiency level of 3 out of 6, which indicated average performance relative to the other countries included. Students in 9 countries performed significantly better, including Shanghai-China, Korea, Finland, Hong Kong-China, Singapore, Canada, New Zealand, Japan, and Australia (OECD, 2009).

In addition to disappointing reading performance, gaps in achievement among specific groups of students reveal that the American education system is consistently failing certain students. In particular, students of color, those from low-income families, and males consistently

underperform relative to White students, those from higher income families, and females. For example, the 2011 NAEP results indicated that 44 percent of White students scored at the proficient or advanced performance reading level, while only 16 percent of Black students assessed demonstrated equivalent performance (NCES, 2011).

The Importance of Early Identification and Remediation

Research over the past several decades suggests that students who do not demonstrate reading success in the early elementary grades will struggle to regain adequate achievement in subsequent years. For instance, Juel (1988) found that students identified as poor readers in first grade were highly likely to remain poor readers in fourth grade, while those identified as proficient readers were significantly more likely to maintain their reading proficiency. A series of subsequent studies have established similar results. In an extension of the work by Juel (1988), researchers found that first-grade reading ability was a significant predictor of reading ability in eleventh grade, even when controlling for cognitive ability (Cunningham & Stanovich, 1997). Furthermore, research on the development of reading has continued to demonstrate the predictive nature of word reading and comprehension skills in early elementary years for later reading ability (Oakhill & Cain, 2012). In particular, kindergarten, first and second grade may represent a salient point for identification and remediation. Although reading ability in first grade is predictive of later reading proficiency, research has shown that students who do not improve by the end of second grade are most likely to experience continued reading difficulties (Spira, Bracken, & Fischel, 2005).

Research indicates that effectively preventing reading difficulties requires early intervention and remediation (Snow, Burns, & Griffin, 1998; Torgesen, 1998). As a result, reading intervention research has been prolific, offering a strong body of evidence for practices

that are effective in remediating reading problems, especially in these early grades. In particular, reading instruction that is focused on developing important core reading skills, gradually introduces more difficult skills, and uses explicit teaching practices with specific feedback and repeated opportunities for practice has been found most effective (Denton, 2012). Meta-analyses suggest that for students who struggle to learn to read, small group interventions that employ these strategies are largely successful in remediating reading difficulties. However, research also suggests that remediation is most successful in kindergarten and first grade (Wanzek & Vaughn, 2007). The importance of effective reading instruction at the early elementary grades is also being emphasized to educators, as can be seen in publications such as the Institute For Education Sciences Practice Guide focused on improving reading comprehension in kindergarten through third grade (Shanahan et al., 2010).

An essential first step in remediating early reading difficulties is identifying those students in need of intervention. Universal screening, the systematic assessment of all students at intervals during the school year, is one method of identifying whether students are making adequate progress toward curricular goals (Vaughn & Fuchs, 2003). Second, universal screening allows educators to identify individual students who are at risk for developing reading difficulties and are, therefore, in need of additional educational supports (Fuchs & Fuchs, 1998). In addition, universal screening contributes to program evaluation decisions, supporting educators' judgments about whether the group (e.g., classroom, school) is achieving proficiency, with inadequate achievement suggesting a need for instructional changes at the group level.

Universal screening, with its potential for identifying students for effective early reading intervention, has become a widespread practice in schools across the country (Fuchs & Vaughn, 2012). One commonly used set of universal screening measures in kindergarten through third

grade, for example, is the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2011). Currently, for first graders, these measures include assessments of students' ability to rapidly name letters, segment words into individual phonemes, decode nonsense words, and read connected text aloud. Unfortunately, existing screening measures such as first grade DIBELS assessments, demonstrate a host of problems. These problems can be organized into two main categories: scientific issues and practice issues. Scientific issues are primarily those concerned with the psychometric properties of the measures, while practice issues include problems with the efficiency and acceptability of current assessments.

The purpose of this chapter is to articulate these two primary issues with current screening approaches in the early elementary grades, especially kindergarten and first grade. First, a historical overview of the development of current reading universal screening measures is presented, as well as evidence for the most popular and psychometrically supported screening measure, Reading CBM (R-CBM; Ball & Christ, 2012). Then, each of the problem areas associated with current screening assessments is presented. Finally, this chapter concludes with an introduction to a newly developed screening tool, Highly Decodable Passages (HD passages; Shinn, 2009; 2012), which attempts to offer a solution to problems with existing measures; research questions intended to investigate the scientific and practical aspects of this measure are outlined.

Curriculum-Based Measurement in Universal Screening

Researchers have sought to identify skills that can be briefly assessed using measures that are both predictive of later reading ability and sensitive to growth over time, allowing for frequent progress monitoring and formative evaluation. In particular, researchers have attempted

to identify developmentally appropriate skills for measurement that can be directly linked to instructional decision making by drawing from students' curricular goals (Shapiro, 2011).

In an effort to combine the advantages of standardized tests and informal teacher observation, while making assessment results relevant to instruction, Stan Deno, along with colleagues at the University of Minnesota, developed an approach to measurement known as curriculum-based measurement (CBM; Deno, 1985; 1992; 1993). The aim of early development research was to identify tasks that could be used as indicators of student achievement for frequent progress monitoring for students with severe achievement discrepancies, including students with IEPs, were simple and efficient to administer, were easy to interpret, and were inexpensive.

Among a number of potential measures that could be used to monitor student progress, researchers investigated a cloze task (identifying an appropriate word to replace a deleted word in a passage), a word meaning task (articulating the meaning of an underlined word in a passage), and reading grade level text aloud. Early research indicated high correlations between two of these procedures (cloze task and reading aloud) and generally accepted criterion measures, such as subtests from the Stanford Diagnostic Reading Test (SAT; Karlsen, Madden, & Gardner, 1975) and the Woodcock Reading Mastery Test (Woodcock, 1973). In particular, the researchers found that reading aloud from text was highly correlated with comprehension measures (Deno, 1985).

Reading CBM

Stemming from the research conducted by Stan Deno (1985; 1992; 1993), Oral Reading, or Reading CBM (R-CBM), represents a hallmark CBM that is currently the most popular method for reading universal screening and progress monitoring (Ball & Christ, 2012).

Beginning in the 1980s, R-CBM became more widely available to educators, opening the door for published measures. In particular, the Test of Oral Reading Fluency (TORF; Children's Educational Services, 1987) emerged in the late 1980s. Other commercially available measures, such as AIMSweb (Edformation, 2005) and DIBELS (1996) emerged later, providing educators with access to a variety of standardized, norm-referenced R-CBMs (Deno & Marston, 2006).

Shinn (1989) described standard R-CBM assessment procedures that were incorporated into universal screening, a process where all students are assessed in order to determine whether they demonstrate performance at a predetermined benchmark that predicts future reading success. In contrast to use of R-CBM for frequent (e.g., weekly) progress monitoring, when screening, students read three grade-level passages aloud for 1 minute. The student's median words read correctly in 1 minute is used as an indication of overall reading proficiency. In addition to being simple to administer and commercially available, R-CBM has developed a substantial literature base and enjoys strong support as a psychometrically sound method for universal screening, particularly in the middle elementary grades (Ball & Christ, 2012).

The National Center on Response to Intervention (NCRTI; http://www.rti4success.org/screeningTools) has produced a tools chart for easily reviewing evidence for a variety of screening measures. The NCRTI established a Technical review that evaluated the scientific rigor of screening tools that were submitted to the committee against an independently established set of criteria. A number of the tools evaluated use R-CBM procedures, such as AIMSweb R-CBM (Pearson, 2012b), DIBELS Next Oral Reading Fluency (DORF; Good & Kaminski, 2011), easyCBM (Alonzo, Tindal, Ulmer, & Glasgow, 2006), and many of these demonstrate convincing or partially convincing evidence of classification accuracy, generalizability, reliability, and validity according to predetermined criteria.

Evidence from the manuals for two published R-CBM measures, DIBELS Next (Good et al., 2013) and AIMSweb (Pearson, 2012a), indicate strong psychometric properties for students in the middle and upper elementary grades. For instance, for DORF alternate form reliability for words read correct using single forms was reported to be .95 at first grade, and ranges from .84 to .95 across first through fifth grade. Using three DORF passages, test-retest ranged from .91 to .97 for Grades 1 through 5 (Good et al., 2013). For AIMSweb oral reading measures, alternate form reliability was reported to be high, with means for each grade level ranging from .93 to .95 (Pearson, 2012a). Externally conducted studies of oral reading measures offer additional evidence for reliability as a screening measure. For example, a review of evidence for DIBELS measures indicated that DORF demonstrated test-retest, alternate form and inter-rater reliability coefficients ranging from .82 to .93 in the literature (Goffreda & DiPerna, 2010).

Additionally, studies have indicated that measures of oral reading are highly correlated with external criterion measures of reading ability. Specifically, correlations between measures of oral reading and concurrently administered standardized, norm-referenced criterion measures of academic achievement, such as the Group Reading Achievement and Diagnostic Evaluation (GRADE; Williams, 2001) and the Metropolitan Achievement Test (Harcourt Brace Educational Measurement, 2000) offer support for concurrent validity (e.g., Goffreda & DiPerna, 2010; Good et al., 2013; Pearson, 2012a). Predictive validity has been established by demonstrating moderate to high correlations with these same types of external criterion measures, as well as state and national achievement tests (e.g., Goffreda & DiPerna, 2010; Goffreda, DiPerna, & Pedersen, 2009; Good et al., 2013; Pearson, 2012a). For example, information provided in the DIBELS Next technical manual indicates that correlations between fall DORF scores (words

read correct) and total scores on the Group Reading Assessment and Diagnostic Evaluation (GRADE; Williams, 2001) range from .64 to .77 in fourth grade (Good et al., 2013).

Given that the purpose of universal screening is to identify students at risk for reading difficulties, diagnostic accuracy, or the ability of a measure to discriminate between students who do and do not demonstrate later academic failure, has become a strong focus of the research on R-CBM. In particular, researchers are interested in the *sensitivity* [the degree to which a measure accurately identifies those students who go on to have difficulties in reading based on a future criterion measure (i.e., true positives)] and *specificity* [the degree to which a screening measure accurately identifies those students who *will not* go on to have reading difficulties (i.e., true negatives)] of potential screening measures (Jenkins, Hudson, & Johnson, 2007). Jenkins, Hudson, and Johnson (2007) asserted that researchers should identify screening measures that demonstrate 90-95% sensitivity and "produce the highest specificity (fewest false positives)" (p. 599). Another measure of diagnostic accuracy is the Receiver Operating Characteristic (ROC) Area Under the Curve (AUC), which is represented by a number between 0.5 (equivalent to chance) and 1.0 (perfect accuracy). Criteria established by Swets (1992) classify AUC values as excellent ($\geq$ .90), good (.80 to .89), fair (.70 to .79), or poor (< .70).

Using DORF scores to predict results on the Pennsylvania state achievement test, Shapiro, Solari, and Petscher (2008) found that sensitivity (correctly identifying students who go on to demonstrate reading difficulties) ranged from .79 to .96 for fall and winter screening assessments in Grades 3 through 5. Specificity, or the measure's ability to correctly identify those students who do *not* go on to demonstrate reading difficulties, ranged form .49 to .61.

The researchers also utilized the AUC (Shapiro, Solari, & Petscher, 2008). Results indicated strong diagnostic accuracy for DORF at these grades, ranging from .87 in the fall,

to .89 in the winter. Other studies examining diagnostic accuracy of R-CBM scores have identified similarly encouraging results when predicting to both commercially available external criterion measures of achievement and state achievement tests (e.g., Keller-Margulis, Shapiro, & Hintze, 2008; Pearson, 2012a; Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008).

Scientific Issues with Current Reading Screening Measures for Early Elementary Students

As stated previously, R-CBM represents a popular approach to universal screening that is efficient, and enjoys strong psychometric support. However, the dynamic nature of reading development would suggest that screening measures that are successful for certain grades may not be effective for all grades. Indeed, although researchers have sought to extend measures of oral reading to the early elementary grades, studies fail to provide strong psychometric support for their use at this level as a method for identifying students in need of academic support. Furthermore, studies of alternative reading screening measures developed for kindergarten and first grade have also failed to offer convincing evidence of their psychometric strength. These issues are explored in greater detail in the following section.

Scientific issues with Reading CBM. Although R-CBM is well established for older elementary students, oral reading measures lack utility for identifying at risk kindergarten and Grade 1 students. A primary concern is that even in the middle of Grade 1, too many students fail to perform well on graded passages. For example, after examining the frequency histograms of DIBELS screening assessments for over 18,000 students followed from kindergarten through grade 1, Catts, Petscher, Schatschneider, Bridges, and Mendoza (2009) found substantial floor effects for DORF in first grade. Specifically, scores were clustered largely at the low end of the score distribution, indicating many students lacked the skills to read more than just a few words

from first grade DORF passages. Histograms did not appear to begin to normalize until at least the middle to end of second grade.

Other studies investigating diagnostic accuracy of oral reading measures in first grade have also raised questions as to whether this approach is a valid method for universal screening at this age. Specifically, one group of researchers found that measures of oral reading offered essentially no more accurate classification than if educators had assumed all students would go on to be successful readers (Johnson, Jenkins, Petscher, & Catts, 2009). Other studies have consistently found that R-CBM, even as late as first grade, demonstrates unacceptable levels of diagnostic accuracy on its own as a screening measure (e.g., Goffreda, DiPerna, & Pedersen, 2009; Jenkins, Hudson, & Johnson, 2007).

Scientific issues with other reading screening measures. Attempts to extend CBM to very early readers in kindergarten and first grade were first established by Kaminski (1992), which eventually became known as DIBELS. In her early work as a doctoral student, Kaminski (1992) identified Letter Naming Fluency (LNF; number of letters named correctly in one minute), Picture Naming Fluency (PNF; number of pictures named correctly in one minute), and Phonemic Segmentation Fluency (PSF; number of phonemes correctly produced in one minute) as potential screening tools for early readers. Correlations with external measures of reading achievement [(e.g., R-CBM, Stanford Diagnostic Reading Test (Karlsen & Gardner, 1985)] ranged from moderate to high for a group of kindergarten ($N = 37$) students ($p < .01$). However, for first grade ($N = 41$) students, only correlations between some of the criterion measures and two of the screening measures (LNF and PNF) were significant ($p < .05$). Kaminski (1992) concluded that LNF, PNF, and PSF were reliable and valid reading screening measures for kindergarten students, but that in first grade, only LNF demonstrated technical adequacy.

Furthermore, Kaminski and Good (1996) suggested that in first grade, R-CBM would be appropriate given students' more advanced reading skill at this age.

Since becoming a publicly available measurement system additions have been made to the DIBELS measures recommended for kindergarten and first grade. The DIBELS Next (Good & Kaminski, 2011) recommends that students in kindergarten be screened using LNF at the beginning, middle, and end of kindergarten, while PSF should be administered in the middle and end of kindergarten, as well as at the beginning of first grade. Two new measures First Sound Fluency (FSF; number of initial phonemes in words produced in one minute), which is administered in the fall and winter of kindergarten, and Nonsense Word Fluency (NWF; number of nonsense words read correctly in one minute), which is administered from the middle of kindergarten until the beginning of second grade, have been added since Kaminski's (1992) original research. DORF is administered to students as a reading screening measure beginning in the middle of first grade and continuing through sixth grade.

Unfortunately, given the research described previously, R-CBM is clearly not a technically adequate screening measure for first graders. Additionally, the developing research on alternative early literacy screening measures, such as LNF and PSF, has failed to demonstrate strong psychometric properties that should be expected for use during such a vital window of reading development as kindergarten and first grade. This creates a huge gap in reading screening literature and practice that must be addressed.

The DIBELS measures LNF, PSF, and NWF are all associated with floor effects in either kindergarten or first grade when they are first administered. For first-grade students, in addition to DORF, NWF demonstrates floor effects throughout the year (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009).

Additionally, Johnson, Jenkins, Petscher, and Catts (2009) found that screening measures other than DORF [Initial Sound Fluency (ISF; earlier version of FSF), PSF, and NWF] produced poor diagnostic accuracy when predicting outcomes on a standardized measure of reading ability administered at the end of first grade. Specifically, the authors investigated the accuracy with which DIBELS measures administered at the end of kindergarten and beginning of first grade (N = 12,055) predicted whether a student would go on to score above or below the 40^{th} or 20^{th} percentile on the SAT-10 (Harcourt, 2003) at the end of first grade. Furthermore, the researchers considered the low base rate of reading difficulty in the population they studied by comparing the classification accuracy of using DIBELS measures versus using no screening measure and assuming no students would go on to experience reading difficulties. Results indicated that had the schools chosen to skip screening altogether, 71.4% of all students would have been correctly classified (as not at risk). When using the most effective of the DIBELS screening measures for kindergarten (NWF) and first grade (DORF), classification accuracy increased by just 4% and 5.5% respectively, raising the question as to whether these screening measures are of any value to educators at all. Even more astonishing, is that rather than using DIBELS designated cut points for predicting risk status, the researchers used statistically optimal cut points. Even when cut points determined statistically to yield the highest classification accuracy, it was still only marginally better than chance.

In a different study by Clemens, Hilt-Panahon, Shapiro, and Yoon (2012) researchers found that the DIBELS measures ISF, LNF, PSF, and NWF, were *not* strong indicators of later reading proficiency. The researchers followed students (N = 101) from kindergarten through first grade, collecting universal screening data using DIBELS measures at the beginning, middle, and end of each year. Using DORF performance in the spring of first grade, For instance, ISF

and PSF demonstrated average AUC values of .698 and .645 respectively, indicating that these measures incorrectly identified a large number of students who would go on to demonstrate oral reading difficulties in the spring of first grade as not at risk in kindergarten and first grade. Average AUC values for LNF and NWF were both over .80. However, researchers observed a high rate of growth on the NWF measure from winter to spring of first grade for students who would later go on to perform below the 30th percentile. They noted that as a result of this sharp increase, educators may inaccurately conclude that struggling readers are improving enough to avoid falling below a later benchmark (Clemens, Hilt-Panahon, Shapiro, & Yoon, 2012).

In addition to data indicating psychometric limitations of existing reading screening measures with early elementary students, the issue of classroom variability has been largely neglected, and represents a critical gap in the research. Although traditional investigations of predictive validity have been frequently employed, such as correlating measures with external criterion measures, and diagnostic accuracy analyses have become increasingly common, investigations of how classroom variability affects predictive validity are sorely needed. In particular, researchers must aim to understand the extent to which student scores on screening measures vary based on classroom membership, and how this impacts the relation between assessments at various time points.

This is a critical point in examining potential screening measures because of the unique structure of schools. As a result of the fact that students are instructed in specific groups (classrooms) by different teachers, it is reasonable to suspect that researchers might observe variability between classrooms in terms of performance on screening measures. Analyses such as hierarchical linear modeling (HLM), which account for nested data, such as students within

classrooms, are necessary for determining whether this variability exists, and how it affects the utility of reading screening measures.

Practical Issues with Current Reading Screening Measures for Early Elementary Students

In addition to scientific issues with current early reading screening measures, such measures also present problems for educators in terms of the efficiency with which they can be applied in schools and their acceptability to educators. Problems such as the amount of time spent testing students with multiple measures, and objections that current early literacy CBMs encourage "teaching to the test" pose serious issues for existing reading screening measures.

Efficiency issues. Currently, for example, DIBELS authors recommend assessing students in kindergarten and first grade with up to four separate assessments (Good & Kaminski, 2011). In the middle of kindergarten, for example, schools using the DIBELS to screen for students who are at risk for reading difficulties are advised to administer FSF, LNF, PSF, and NWF. In the beginning of first grade, schools are advised to administer LNF, PSF, and NWF as well. These multiple assessments, at multiple time points throughout the year, require school staff to spend valuable time assessing students.

Recent studies raise the question as to whether the use of multiple assessments is actually worth the time spent administering them. In the study by Johnson, Jenkins, Petscher, and Catts (2009), described previously, the authors examined the added value of additional DIBELS screening measures over the highest performing measure alone in predicting performance on the SAT-10. When predicting performance below the 40th percentile at the end of first grade from the beginning of first grade, combining DORF scores with the DIBELS measure with the next highest associated specificity (NWF; 42% when setting sensitivity to 90%) resulted in an improvement in specificity of less than 1%. When predicting performance below the 20th

percentile and maintaining 90% sensitivity, specificity increased from 65% to 67% when combining NWF with DORF. Such marginal improvements are concerning in light of how much time added assessments take to administer. While assessments take only one minute each to administer to students (3 minutes for 3 R-CBM probes during screening assessments), a few additional minutes add up quickly when assessing an entire district.

Indeed, in a survey of researchers' and educators' perspectives of DIBELS assessments some disadvantages of current measures cited by both survey respondents and individuals in follow-up interviews included the time spent testing and the use of nonsense words (Hoffman, Jenkins, & Dunlap, 2009). With respect to measures in kindergarten and first grade this is likely to be a very serious issue, because students in these grades require a greater number of early literacy assessments. This may be less of a concern in the later elementary grades, when oral reading measures alone are effective for screening (e.g., Ball & Christ, 2012; Keller-Margulis, Shapiro, & Hintze, 2008; Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008).

Acceptability issues. Screening measures for early elementary students present another problem because of issues with authenticity and acceptability. In reaction to an article by Riedel (2007), for instance, Samuels (2007) raised the question as to whether the DIBELS LNF, PSF, and NWF could even be justifiably applied in schools as measures of reading ability. Specifically, Samuels questioned whether speed with such specific skills was truly indicative of reading proficiency. Others have noted that measures of the rate at which a student can demonstrate a specific skill, such as naming letters, are not authentic assessments of reading ability. That is, timed assessments of specific skills like letter knowledge and phonemic awareness fail to measure what Pearson (2006) refers to as "real reading," which includes comprehension and critical thinking.

The concept of authenticity in assessment has become particularly important in early childhood. For instance, in his book *Authentic Assessment for Early Childhood Intervention: Best Practices*, Bagnato (2007) emphasizes that fact that "contrived tasks and materials" included in conventional assessments are inadequate forms of assessment compared to observations of the child in his/her natural environment. Furthermore, he argues that psychometric test items are too far removed from the curriculum, are administered by those who are unfamiliar with the child, and encourage teaching to the test. Conversely, authentic assessment urges educators to understand a student's level within his/her specific curriculum and invites adults who are intimately familiar with the child's development to be involved in the assessment process.

These objections to conventional assessment in early childhood closely mirror objections raised against the DIBELS, assessments intended for students just a few years older. In addition to arguments that the DIBELS assessments reduce reading to a few component reading skills, Goodman (2006) also asserts that DIBELS assessments require students to demonstrate many of these skills out of context. He argues that it would be more logical to assess phonemic awareness, for instance, in the context of spoken language, rather than asking students to abstract these skills to contrived assessment procedures included in tests like ISF.

Another major objection to DIBELS assessments is the perspective that teachers end up teaching to the test, with DIBELS assessments guiding curriculum choices at the expense of quality reading instruction (Pearson, 2006; Goodman, 2006). For instance, Goodman (2006) asks if children who fail to meet NWF benchmark scores should be taught to read words based on their sounds alone, ignoring context cues, regardless of how silly the words sound. Additionally, just as Bagnato emphasizes the need to involve adults who know the child well in

early childhood assessment, Goodman (2006) asserts that teachers should be allowed the flexibility to use their professional judgment in assessing and instructing their students.

Possible Alternatives to Current Reading Screening Measures for Early Elementary Students

Given the scientific (e.g., psychometric weaknesses of some measures) and practical (e.g., problems with efficiency and acceptability) issues with current reading screening measures for early elementary students researchers have attempted to identify alternative measures.

Word identification fluency. One approach that has received attention is word identification fluency (WIF; Fuchs, 2003), where a random selection of high frequency words from the Dolch list is presented in isolation, rather than in the context of connected text. Like other oral reading CBM measures, students read these words aloud for 1 minute and the number correct are counted. Fuchs (2003) demonstrated that in the middle of first grade, educators can use a benchmark of 40 words read correct in 1 minute to predict future reading success.

In a comparison with the DIBELS NWF, Fuchs, Fuchs, and Compton (2004) found that the WIF was a more valid assessment for predicting end of year reading outcomes in first grade. Specifically, the authors assessed 151 first graders at risk for reading difficulties in the fall and spring using WIF and NWF. Word Identification and Word Attack subtests from the Woodcock Reading Mastery Test (Woodcock, 1987) were administered to students at both time points, while the Comprehensive Reading Assessment Battery (CRAB; Fuchs, Fuchs, & Hamlett, 1989) was administered in the spring only. Results indicated that WIF was statistically superior to NWF in terms of concurrent and predictive validity. Correlations between WIF and the Word Identification measure were significantly higher than for NWF ($p < .001$) in the fall and spring. Additionally, both the CRAB Fluency and Comprehension measures administered in the spring

demonstrated significantly higher correlations with WIF than with NWF ($p < .01$). In a direct comparison of the predictive validity of NWF and WIF using dominance analysis, the authors found that WIF fall scores and slope across the year was statistically superior to NWF in predicting spring outcome measures.

In another study that included measures of LNF, PSF and NWF, researchers found that WIF tended to demonstrate the highest classification accuracy for first grade readers (Clemens, Shapiro, & Thoemmes, 2011). The study included 138 first grade students who were assessed in the fall using WIF, LNF, PSF, and NWF. In the spring, students were assessed using DORF, the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999), and AIMSweb Maze. Logistic regression analyses indicated that only WIF was a significant predictor of TOWRE subtests, AIMSweb Maze, and a composite of spring TOWRE, Maze, and DORF assessments ($p < .01$). WIF and PSF were both significant predictors of spring DORF scores alone ($p < .01$).

When analyzing AUC values, WIF also demonstrated the highest overall classification accuracy, with values ranging from .86 to .91 when predicting performance below the 30th percentile on spring outcome measures (Clemens, Shapiro, & Thoemmes, 2011). When the sensitivity was held constant at .90, WIF and LNF demonstrated the highest rates of specificity (.52-.71 and .56-.69 respectively). AUC values and specificity were not consistently improved as a result of adding a DIBELS screening measure to WIF alone. According to the authors, WIF and PSF together appeared to represent the most parsimonious set of screening measures for achieving the highest rate of classification accuracy.

Decodable text. The limited research to date suggests that WIF may improve on the psychometrics of early reading screening compared to current measures, such as DIBELS LNF,

PSF, and NWF. Additionally, WIF offers the potential benefits of reducing the practice concerns of efficiency and acceptability. Specifically, the use of just one measure would reduce testing time required for the administration of 3 to 4 measures for every student. Additionally, some of the issues regarding authentic assessment *may* be reduced by asking students to read real English words in isolation, rather than reading nonsense words, or demonstrating phonemic awareness through decontextualized tasks like articulating sounds in words or naming letters.

Unfortunately, isolated word reading measures are unlikely to address all of the authenticity concerns raised by those who feel many curriculum-based measures do not fairly assess a student's ability to read connected text. WIF may present another problem in that high frequency words, such as those from the Dolch list, may not fully represent the components of phonics, which should be addressed in early elementary literacy instruction. Additionally, in their original work comparing WIF to fluency reading connected text, Deno, Mirkin and Chiang (1982) found that oral reading demonstrated stronger psychometric properties than word reading measures. However, given the limitations of current Grade 1 R-CBM measures, creating developmentally appropriate passages that are capable of validly evaluating a student's reading ability is a key hurdle to overcome.

These issues are especially evident in the beginning of first grade. While some early literacy measures, such as those developed by Kaminski (1992; e.g., LNF, PSF) may offer more reliable predictions of student performance in kindergarten, these measures demonstrate limitations in first grade, as Kaminski pointed out in her dissertation research and subsequent research has reiterated (e.g., Catts et al., 2009; Clemens, Shapiro, & Thoemmes, 2011). However, while Kaminski (1992) asserted that R-CBM would be appropriate for first-grade students because of their emerging reading abilities, the research outlined here clearly indicates

that this is not the case. As a result, there is a significant gap in the measurement literature in terms of identifying assessments that are both psychometrically strong and acceptable to school practitioners for screening early readers in the beginning of first grade.

Passages comprised of decodable text may offer an opportunity to assess early readers' fluency with connected text in a way that ameliorates these concerns. Decodable text is considered text that includes a high proportion of words that are phonetically regular, with a large number of the letter-sound relationships included in the text being ones that the student has learned (e.g., Mesmer, 2001). Therefore, in contrast to WIF, decodable text passages should not only provide students an opportunity to demonstrate their ability to recognize common high frequency words, but also to use phonics skills to decode developmentally appropriate words, and to demonstrate features of good reading with connected text. Possessing both the scientific advantages of WIF, as well as the authenticity and efficiency benefits of R-CBM would position decodable text as a viable alternative to current reading screening measures for beginning first grade, in particular.

There is evidence to suggest that when students are assessed using decodable text, it provides a better indication of the extent of their phonics knowledge. In a study by Mesmer (2005), the author found that when first-grade students receiving a phonics intervention were assessed using decodable text, they demonstrated better word attack skills than students who received the same intervention but were assessed using less decodable text. Other research has indicated that having difficulty decoding words affects early elementary students' reading speed, accuracy, and comprehension (e.g., Hiebert & Fisher, 2007). Results of this nature have led researchers to encourage the consideration of text leveling factors, such as the proportion of

decodable words, in the development of reading assessments for younger elementary students (Paris & Hoffman, 2004).

Additionally, research conducted by Lopez, Thompson, and Walker-Dalhouse (2011), suggests that proficient readers rely on the context of written passages (i.e., context of additional words in connected text) to improve reading fluency, whereas less proficient readers do not. Results such as these serve to underscore the importance of utilizing connected text as an assessment of reading proficiency, rather than isolated word lists or individual subskill assessments. Furthermore, initial research conducted by Shinn (2009; 2012), indicates that highly decodable passages offer a viable method of universal screening with early elementary students. However, current screening measures have not utilized decodable text, and published research has not investigated the extent to which such assessments effectively identify students at risk of future reading difficulties. This represents a serious gap in the literature, as well as an opportunity to develop measures for young readers that might be more acceptable to consumers.

Highly Decodable Passages

In response to concerns over current early reading screening and progress monitoring measures, Shinn (2009; 2012) has developed a set of highly decodable passages (HD passages), designed for potential use as screening and progress monitoring measures in kindergarten and first grade, as students are developing early reading skills. These passages are intended to capitalize on the advantages of both word identification fluency and oral reading measures by including a high proportion of phonetically readable words, as well as developmentally appropriate sight words that are organized into connected text passages. Like other R-CBM measures, the metric used to assess student reading ability with these passages is words read correct per minute. However, the passages are intended to be more appropriate to the

development of the reading process for younger students, in late kindergarten and first grade, who are beginning to learn and successfully apply phonological decoding skills (Spear-Swerling, 2004).

Given the current gaps in the literature on universal screening for reading in early elementary school, the purpose of this study is to evaluate the psychometric properties of these HD passages. Specifically, the focus of the study is to investigate the utility of these passages as a screening measure for first-grade students with potential improvements in science, and practice. To achieve this goal, the study attempted to answer with following research questions:

1. What is the reliability (i.e., inter-rater, test-retest, and alternate form reliability) of HD passages?
2. What is the convergent validity (i.e., concurrent and predictive validity) of HD passages for assessing reading ability in Grade 1 with currently available measures of reading ability (e.g., DIBELS Next NWF, DORF, GRADE)?
3. How well do students' winter HD passage scores predict their spring HD passage scores, and how does inter-classroom variability affect this relation?
4. What is the winter to spring diagnostic accuracy (i.e., sensitivity, specificity, negative & positive predictive power, negative and positive likelihood ratios, post-test probabilities) of HD passages when predicting to currently available measures of reading achievement (e.g., DIBELS Next NWF, DORF, GRADE)?
5. How does teacher acceptability of highly decodable passages compare to acceptability of currently available Grade 1 reading screening measures?

Chapter II: Review of the Literature

Universal screening represents a well researched and widely adopted practice in American schools. Educators and researchers recognize the importance of identifying students at risk of future reading difficulties as early as possible in order to provide effective intervention. Although schools throughout the United States have drastically increased the frequency with which universal screening measures are used, there continue to be significant limitations to the measures utilized with students in the early elementary grades. In particular, these measures continue to demonstrate problems with predictive utility, failing to identify students who will go on to experience reading difficulties with the accuracy that is expected (Fuchs & Vaughn, 2012).

As outlined in Chapter I, the purpose of the current study is to investigate the utility of a newly developed universal screening measure, known as HD passages. In order to better understand the potential this measure has as a screening measure in the early elementary grades, the current chapter will outline pertinent areas of research. First, this chapter will present literature on the theory and research on reading development, followed by a discussion of the development of curriculum-based measurement, R-CBM – a commonly used universal screening measure in elementary schools (Ball & Christ, 2012) – and early literacy CBMs. Next, the chapter will outline limitations of current screening measures for early elementary readers with a focus on psychometric issues. Finally, greater detail regarding possible alternatives to current curriculum-based measures will be discussed.

Reading Processes and Reading Development

Theories of Reading

Several theories of reading emphasize the importance of developing automaticity with reading subskills in order to engage in higher-level reading processes, such as comprehension of

connected text. One example of such a theory is the LaBerge-Samuels model of processing in reading (LaBerge & Samuels, 1974). In their model, LaBerge and Samuels emphasize the fact that many processes must be coordinated in order for reading comprehension to occur. Furthermore, they point out that for most adult readers, this happens very quickly. In order for such a complex process to occur so quickly, many processes must be automatic, in order for attention to be diverted to more difficult processes. This is based upon the premise that while we can actively attend to very few things at once, most likely only one, when processes have been automatized we can perform many at once.

According to the LaBerge-Samuels (1974) model of reading, attention regulates a series of processes. The first of these is visual perception of features of text, which requires readers to access visual memory. Once features of text are received in the visual memory, phonological memory is accessed to understand what words are represented by written text. Readers can draw upon episodic memory when reading novel words, in order to contextualize the newly learned word in memory. Finally, the meanings of words and comprehension of connected text messages is retrieved from semantic memory. This model outlines different processes that readers may need to devote more or less time to, depending on their skill with reading.

A very early reader may utilize a great deal of attention to focus on and identify individual letters, spending time identifying them, and then assigning them sounds, which can finally be blended into words (LaBerge & Samuels, 1974). After exerting so many resources to decode an individual word, the LaBerge-Samuels model argues that not much will be left in terms of cognitive resources for comprehension. Furthermore, this process will have taken a great deal of time to complete, whereas more proficient readers will be able to automatically

recognize letters and words on a page, and to retrieve necessary information from semantic memory while organizing that information in a way that makes sense of the text's message.

The importance of developing automaticity with reading skills in order to increase reading fluency is further underscored by the work of Just and Carpenter (1980). In their model of reading, eye fixation provides the opportunity to learn about the reading process. According to the researchers, readers modulate the rate of input from text in order to match the rate at which they are able to comprehend that input. Following this logic, time spent gazing at a word can be considered a measure of an individual's processing time necessary for reading comprehension.

The model proposed by Just and Carpenter (1980) explains this reasoning by proposing that the reading process begins by exacting visual input from an individual word, which is then identified and encoded, assigned meaning, and integrated with previous word knowledge in a sentence. Working memory holds and provides information to facilitate this process, while drawing on long-term memory to understand the text. The reader's gaze can then be moved to the next word when this process has been completed. In research conducted by Just and Carpenter, the authors found evidence for their proposed model. Specifically, they found that college students' actual gaze durations (in milliseconds) for individual words in passages closely matched estimations generated based on their model.

Similarly, Perfetti's (1985) verbal efficiency theory attempts to explain individual differences in reading comprehension by illustrating the role of working memory in the reading process. According to the theory, differences in reading comprehension are primarily a result of the efficiency with which an individual can represent text in working memory. Specifically, the more efficiently an individual can receive, code, and integrate propositions (smallest units of meaning) from text in working memory, the more effectively he or she will be able to

comprehend the text. This theory also acknowledges that differences in long-term memory will contribute to reading comprehension as well, such as individual schemata, by recognizing that information must be retrieved from long-term memory into working memory in order for text processing to occur.

In contrast to Just and Carpenter's (1980) model of reading, Perfetti's verbal efficiency theory asserts that individuals are not capable of independently regulating the input they receive or the cognitive resources they expend in order to do so. Instead, resources are allocated according to how well developed the individual's reading processes are. These processes include decoding/identifying a word, assigning meaning to the word or phrase, and comprehending text. Reading is considered efficient when the outcomes of reading processes are of high quality, while the cognitive resources expended remain low.

Verbal efficiency theory (Perfetti, 1985) identifies word identification, or what Perfetti terms "lexical access," the most likely process to be made very efficient, or automatic. Additionally, some aspects of propositional encoding, which require schema activation, might also be made efficient, according to the theory. When readers have achieved efficiency with these processes through repeated practice and learning, this leaves resources available in working memory to encode and integrate propositions, to make inferences, and to interpret and critically comprehend the text being read. This aspect of the verbal efficiency theory is associated closely with the LaBerge-Samuels (1974) model of reading. Both of these perspectives emphasize the limited cognitive resources available to process text, making it essential that lower-level processes, such as word identification, occur with relatively little effort in order for effective comprehension to occur.

Distilling the reading process further, the aptly named Simple View of Reading (SVR; Gough & Tunmer, 1986) considers reading comprehension to be the product of decoding ability (free of context) and listening comprehension. Furthermore, the authors assert that reading difficulties are a result of either decoding or listening comprehension deficits, or a combination of both. According to Hoover and Gough (1990) the notion that reading is a complex process has been falsely embraced by researchers and educators alike.

Results of a longitudinal study conducted by Hoover and Gough (1990) provide evidence for the SVR. The authors followed bilingual students from first through fourth grade, assessing each student annually using measures of nonsense word reading, listening comprehension, and reading comprehension. Using hierarchical multiple regression the authors demonstrated that decoding ability and listening comprehension explained a significant ($p < 0.001$) amount of the variance in reading comprehension ability when combined linearly. However, the product of these two factors explained significantly more of the variance ($p < 0.01$).

In a different study of 4- and 6-year-old English speaking students from the United States and Canada ($N = 232$), researchers also found evidence for the SVR using exploratory factor analysis (Kendeou, Savage, & van den Broek, 2009). Specifically, students were assessed using a set of listening and video viewing comprehension measures. Measures of phonological processing, vocabulary, letter identification, and word identification were also included. Exploratory factor analysis using principal component analysis (PCA) yielded a 2-factor solution, which included items relating to Decoding Skills (phonological awareness, letter identification, vocabulary) and Comprehension Skills (listening comprehension and video viewing comprehension). These same results were replicated using a different set of measures, where the Decoding Skills factor was comprised of DIBELS NWF, ORF, and Retell Fluency, while the

Comprehension Skills factor was comprised of the GRADE Listening Comprehension and DIBELS Story Retell assessments.

Some researchers have suggested modifying the SVR theory somewhat. For instance, Joshi and Aaron (2000) have argued for adopting processing speed as an additive component to the traditional SVR model (decoding x listening comprehension + speed of processing = reading comprehension). In their own research, Joshi and Aaron noted that the inclusion of a measure of processing speed (rapid letter naming) improved the SVR's ability to explain variance in reading comprehension abilities of 40 third grade students. Findings such as these continue to demonstrate the importance of proficiency with basic reading skills in order to successfully comprehend connected text.

Stages of Reading Development

Considering the importance of automaticity or efficiency of basic reading skills, such as decoding, to comprehension, it is essential to understand when a typically developing reader can be expected to demonstrate specific skills, and at what point they are mastered. This understanding is essential in developing universal screening measures that are effective in targeting salient skills at the relevant age or grade level. Ehri and McCormick (1998) provide a map for reading development in their phases of word learning. These stages stress the importance of efficient decoding and the development of sight word knowledge in order for comprehension to occur, in line with models of reading like those outlined previously. A total of 5 phases are included, each of which represents a different level of understanding that the reader has of the alphabetic system.

The first of Ehri and McCormick's (1998) five phases is referred to as the pre-alphabetic phase, and for typically developing children refers to preschool-aged students. At this phase in

the development of reading, children do not use alphabetic knowledge, but instead rely on certain cues and context to identify words. For example, the consistent placement of a word on a familiar toy label may trigger the child's memory for that particular word because of its context, shape, color, and so on. Around kindergarten, students move to the partial-alphabetic phase, at which point they begin to use memory of individual letters to aid their word recognition. Although children in this phase of reading development can associate letters and sounds to some degree, this ability is more fully developed in the full-alphabetic phase. During this phase, children possess strong phonemic awareness and alphabetic knowledge, and are able to match appropriate sounds to printed symbols. Early decoding skills are evident at this stage, allowing for deliberate decoding of phonetically regular words, while a growth in sight word vocabulary is also apparent. The full-alphabetic phase characterizes most typically developing readers in first grade and beyond.

Most students will then transition to the consolidated-alphabetic phase in second grade (Ehri & McCormick, 1998). This is indicated by students' ability to recognize spelling patterns in words, and to associate those patterns with their corresponding sounds in order to decode more complex words. Readers in this phase expand their skills rapidly by learning essential parts of language, such as suffixes and prefixes, developing a larger sight word vocabulary, and understanding linguistic rules that govern word pronunciations. Eventually, readers will transition into the automatic phase, which is characterized by a very large sight word vocabulary, and efficient word-attack skills. Reading becomes highly efficient at this point because readers encounter mostly sight words when reading, or can quickly identify a word using numerous strategies that have been well developed. As outlined in models and theories of reading

described previously, this allows for more cognitive resources, namely attention, to be devoted to comprehension (Just & Carpenter, 1980: LaBerge & Samuels, 1974; Perfetti, 1985).

Developmental phases of reading have also been established by Spear-Swerling and Sternberg (1994), who provide a greater focus on comprehension in their phases. The first two phases (visual-cue recognition phase and phonetic-cue word recognition phase) are very similar to those included in the reading development process described by Ehri and McCormick (1998). Specifically, typically developing readers often demonstrate limited alphabetic knowledge and a reliance on non-text visual cues in preschool, followed by early knowledge of letter-sound correspondences and developing phonemic awareness. Spear-Swerling and Sternberg (1994) add that children transition from having proficient oral language comprehension only, to very early text comprehension as well.

Subsequent phases of Spear-Swerling and Sternberg's (1994) reading development process (controlled word recognition phase, automatic word recognition phase, strategic reading phase, proficient reading phase) also share overlap with Ehri and McCormick's (1998) phases of reading development, especially in terms of word recognition skills. In terms of comprehension, however, Spear-Swerling and Sternberg outline important specifics about development relative to age and word recognition skill. For instance, especially during the controlled word recognition phase (approximately first to second grade), children's reading comprehension is still very basic, in line with theories that establish comprehension is difficult in the face of strenuous word recognition (e.g., LaBerge & Samuels, 1974). As students transition to about second or third grade, and to the automatic word recognition phase, constraints on comprehension are assumed to come mostly from background knowledge, vocabulary, and knowledge of

comprehension strategies. Eventually, reading comprehension becomes comparable to oral language comprehension.

Empirical Support for the Importance of Automaticity

A number of studies provide empirical support for theoretical assertions that the development of automaticity with lower-level reading skills is essential for higher level reading skills to be effective. For example, in their study on the impact of passage difficulty on student reading fluency, Ardoin, Suldo, Witt, Aldrich, and McDonald (2005) investigated how oral reading fluency was related to various methods of determining text readability. Participants included 99 third grade students in general education. Researchers developed a series of reading probes taken from third and fourth grade reading curriculum, which were administered to students using standard curriculum-based measurement procedures. A total of six probes were evaluated for readability according to eight different procedures (Spache, Fry's graph, Dale-Chall, FOG, Powers-Sumner-Kearl, Flesch-Kincaid, Forecast, and SMOG readability estimates).

To evaluate the impact of text difficulty on reading fluency, the researchers obtained correlations between the eight readability estimates for each of the six probes and subjects' words read correct per minute (Ardoin et al., 2005). Moderate correlations indicated that readability estimates were fairly accurate in predicting the rate at which students could read text, with more difficult passages resulting in lower words read correct per minute. Furthermore, the Forecast and FOG readability estimates were consistently most predictive of the words read correct than the other six readability estimates. Conversely, Spache and Dale-Chall estimates were consistently worse than the others. Additionally, the strategies for determining text difficulty that were most predictive of reading fluency rates were average number of syllables per 100 words and words from the Dale-Chall 3,000 word list. These findings provide support for the fact that oral reading fluency serves as an indicator

of the difficulty of text for a student, with easier texts being read more fluently, and more difficult texts taking longer to decode.

In a similar evaluation of text-leveling procedures, Compton, Appleton, and Hosp (2004) evaluated the impact of various text leveling systems on second grade students' reading fluency and accuracy. The researchers assigned 248 second grade students to either a low decoding ability or average decoding ability category using the Word Attack subtest of the Woodcock Reading Mastery Test-Revised. The sample included a distribution of racial groups and genders, with more minority students and males in the low decoding ability group.

The researchers assessed each student using the Word Identification and Word Attack subtests of the Woodcock Reading Mastery Tests-Revised, as well as weekly R-CBM passages (Compton, Appleton, & Hosp, 2004). Oral reading passages were scored for both words read correct per minute (fluency) and percent of words read correctly (accuracy). Passages were evaluated for readability using the Flesch-Kincaid (average number of syllables per word and sentence) and Spache (average sentence and word length) formulas, as well as decodability and percentage of high frequency words.

Using correlational analyses the authors found that conventional readability formulas were not highly related to one another, while percentage of high frequency words was also not significantly correlated with readability estimates. Conversely, the decodability estimates calculated by the researchers were significantly correlated with Flesch-Kincaid estimates of readability, with more decodable, and fewer multisyllabic words being associated with easier readability estimates. Overall, decodability and proportion of high frequency words predicted reading fluency and accuracy. Specifically, students were able to read passages more fluently and with higher accuracy if the passages contained a higher number of decodable and high frequency words.

In a slightly different approach to evaluating text difficulty, Hiebert and Fisher (2007) rated passages for difficulty according to the number of novel words that appeared and compared difficulty levels to students' reading speed, accuracy, and comprehension. Specifically, for each passage the researchers determined the average number of words per 100 that would be difficult for a student based on his or her exposure to high frequency words in the text and the vowel patterns to which they had been exposed. The authors referred to the number of novel or unique words predicted for each text as the critical word factor (CWF). Using this procedure, four texts were assigned to either the low difficulty (CWF equal to 0) or high difficulty (CWF ranged from 20 to 21) category.

The study included 36 participants in first grade who were selected based on their ability to read at least 5 high frequency words from a list they were provided by the researchers (Hiebert & Fisher, 2007). Students read each of the four passages and were given a score for speed (words correct per minute) and accuracy (words correct per 100 words). They were also asked to retell the story and given a score of 0 (no evidence of comprehension) to 4 (full comprehension). Analysis of variance (ANOVA) was then used to evaluate the difference in reading performance between the two levels of text difficulty. As hypothesized, results indicated that easier passages (low CWF) produced higher rates of reading speed, accuracy, and comprehension compared to more difficult passages (high CWF). In most cases, passages rated at the same level of difficulty did not result in differential reading performance. The exception was reading accuracy, which differed significantly between the two passages rated as difficult (high CWF).

Together, these articles demonstrate the importance of developing fluency with basic reading skills, such as knowledge of the alphabetic principle and phonics, in order for proficient reading to occur. In particular, at the early grades, when students are able to decode or recognize a large number of words by sight, they demonstrate more fluent oral reading and better text comprehension. These

studies also suggest that current methods for calculating readability may be inadequate for developing passages that allow educators to gauge student reading proficiency, especially at the earlier grades.

Curriculum-Based Measurement

Researchers over the past several decades have attempted to develop and refine a set of measures that evaluate progress toward general curricular goals, can be quickly and easily administered, and can be used repeatedly to measure student progress over time (e.g., Deno, 1985; Shinn, 1989). These measures are referred to as curriculum-based measures (CBM). As Deno (1992) explained, at the time these measures were being developed (1977 to 1983), teachers and parents were very unclear as to what key indicators could be used to gauge academic growth. According to Deno, a major limitation of standardized tests is that they separate and assess basic skills individually, when the goal of education is often the integrated application of these multiple skills in a fluent fashion. The objective of CBM was to target salient indicators of general academic outcomes, such as reading proficiency or math computation proficiency in a given grade. In addition, drawing from behavioral theory, these researchers attempted to develop a measurement system that emphasized the individual's response to specific instruction. The aim was to allow educators the opportunity to study the impact their instructional changes had, by providing a method of assessment that could be repeatedly administered and graphed.

By the late 1970's CBM researchers were already encouraging the fields of educational and school psychology to adopt alternative procedures to the standardized tests that were so popular (Jenkins, Deno, & Mirkin, 1979). Researchers argued that standardized assessment instruments were inadequate for program planning, monitoring student progress, program outcome evaluation, and even for special education eligibility. They supported their claims with evidence that the level of standardized test content differed from one test to another, and that

there were major inconsistencies in the scores students obtained on these measures. Furthermore, they pointed out that metrics like percentile ranks were not useful in making instructional decisions, because they do not offer a clear picture of the student's specific skill deficits or strengths.

Reading Curriculum-Based Measures

Curriculum-based measures for reading began appearing in the literature in the early 1980's. Deno, Mirkin, and Chiang (1982) presented results of studies on three CBM approaches to reading, including reading words in isolation, reading words in context, reading connected text, identifying missing words from connected text passages (cloze procedure), and identifying word meanings from connected text passages. Rate correct was used as the metric for scoring the assessments. A total of 18 general education and 15 special education students were administered all five assessment types, as well as two standardized outcome measures; these were subtests from the Woodcock Reading Mastery Test (WRMT; Woodcock, 1973) and the Stanford Diagnostic Reading Test (SDRT; Karlsen, Madden, & Gardner, 1975). Results indicated that reading aloud measures, including reading words in isolation and in context, and reading connected text demonstrated superior correlations with standardized measures compared to the cloze and word meaning tasks. Correlations ranged from .73 to .91 for both general education and special education students. Other studies presented in the same article consistently indicated that oral reading tasks, particularly reading aloud from connected text, were highly correlated with student word reading and comprehension performance on standardized tests.

Eventually, Reading Curriculum-Based Measurement (R-CBM), as measured by words read correct per minute, became the standard curriculum-based measure for reading. Shinn (1989) provided standardized procedures for developing these passages from curriculum

materials, administering them, and interpreting results. Specifically, these standardized procedures involve administering 3 passages of at least 250 words to students. The student reads aloud for 1 minute from each passage, and the median number of words read correctly is taken to account for passage variability.

Originally, educators were advised to select passages from curricular materials that was somewhere in the middle of the difficulty range for students in a given grade. Efforts were made to avoid selecting passages that would be too easy for most students, or too difficult for most students, in order to prevent significant floor or ceiling effects (Shinn, 1989). As the research on CBM progressed, however, there was a move away from drawing from individual curricula. Instead, a shift toward developing generic passages occurred, in an effort to minimize the variability of passages developed by various educators from a variety of different curricular materials (Fuchs & Deno, 1994).

Some of the first standardized R-CBM passages were available in the early 2000's. For example, Gary Germann developed AIMSweb, a comprehensive system for universal screening and progress monitoring that included R-CBM passages. Later acquired by Harcourt Assessment (Pearson, 2006), AIMSweb passages continue to be widely used in school districts across the country, as indicated by the thousands of students included in the normative samples for each grade level passage. The development of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002) and easyCBM (Alonzo, Tindal, Ulmer, & Glasgow, 2006) R-CBM passages provided educators with other standardized systems for screening and progress monitoring. The development of these measurement systems relieved teachers of the burden of having to create their own passages, while providing more standardized materials for evaluating student reading ability. Currently, standardized R-CBM passages

represent the most commonly employed universal screening measures, and enjoy substantial support for their reliability and validity as a tool for identifying students at risk for future reading difficulties (Ball & Christ, 2012).

Early Literacy Curriculum-Based Measures

In her dissertation, Kaminski (1992) outlined a need to develop CBM measures that were reliable and valid for identifying at risk readers in kindergarten and first grade, citing issues with prevalent measures that were similar to those presented in the current paper. Specifically, she noted that standardized tests, most commonly used for identifying at risk readers at the time, present limitations because they prevent early intervention by waiting until students demonstrate significant difficulties relative to their peers. Additionally, Kaminski noted the psychometric limitations of R-CBM for kindergarten and first-grade students, including floor effects and limited variability, which indicated a need for alternative assessment procedures.

The three measures under investigation in Kaminski's (1992) dissertation were Letter Naming Fluency (LNF), Picture Naming Fluency (PNF), and Phonemic Segmentation Fluency (PSF). Results of a study with kindergarten ($N = 37$) and first grade ($N = 41$) students suggested that although these measures may demonstrate reliability and validity for identifying at risk kindergarten readers, they appeared to be inadequate for first-graders. Students were administered the potential screening measures in the fall, as well as several criterion outcome measures. The McCarthy Scale of Children's Ability (MSCA; McCarthy, 1972), Metropolitan Reading Test (MRT; Nurss & McGauvran, 1986; kindergarten only), Rhode Island Pupil Identification Scale (RIPIS; Novak, Bonavantura, & Merenda, 1973), Teacher Rating Scale (TRS), R-CBM (first grade only) and the Stanford Diagnostic Reading Test (SDRT; Karlsen & Gardner, 1985; first grade only) were administered to students in the fall. The RIPIS, R-CBM

(first grade only), and TRS were then administered to students again 9 weeks later. Half of the participants in each grade were also assessed weekly using the early literacy CBMs to investigate the measures' potential as progress monitoring tools.

Mean alternate form reliability point estimates (scores from assessments administered at one point in time) in kindergarten was .93 (LNF), .77 (PNF), and .88 (PSF). For first-grade students, alternate form reliability was .83 (LNF), .62 (PNF), and .60 (PSF). In terms of concurrent validity, correlations between early reading CBM assessments and criterion measures using point estimates ranged from .43 (PSF and RIPIS; $p < .01$) to .85 (LNF and TRS; $p < .01$) for kindergarten students, and from .02 (PSF and MSCA) to .50 (LNF and SDRT; $p < .01$) for first-graders. Lower reliability and validity coefficients led Kaminski (1992) to conclude that the measures under investigation (LNF, PNF, and PSF) "did not work as well for first graders" as they did for kindergarteners, and indicate their limited utility as screening measures for this age group (p. 76).

The results of Kaminski's (1992) work were published in an article by Kaminski and Good (1996). In the article, Kaminski and Good first used the name DIBELS to refer to their measurement system for early readers and established that LNF, PNF, and PSF were not appropriate for first grade students who demonstrated reading abilities, in which case they asserted that R-CBM would be suitable. Currently, DIBELS Next (Good & Kaminski, 2011) recommends that students in kindergarten be screened using FSF, LNF, PSF, and NWF. The authors also recommend that first grade students be administered LNF and PSF (despite previous advice against doing so), in addition to NWF and DORF.

Empirical Support for R-CBM

Given recommendations that R-CBM be used to screen first-grade readers for risk, as well as the measure's popularity in schools, it is essential to understand the empirical support for this assessment, especially as it relates to first grade students. Overall, commercially available R-CBMs report high reliability and validity coefficients. For instance, for DIBELS Next ORF (DORF; Good & Kaminski, 2011) alternate form reliability for words read correct using single forms ranges from .84 to .95 across first through fifth grade, while test-retest for triads ranges from .91 to .97 for grades 1 through 5 (Good et al., 2013). R-CBM passages created by AIMSweb (Pearson, 2012b) also report high alternate form reliability, with means for each grade level ranging from .93 to .95 (Pearson, 2012a). In terms of criterion validity, correlations between AIMSweb R-CBM and state tests ranged from .60 to .72, while correlations between DORF and total scores on the Group Reading Assessment and Diagnostic Evaluation (GRADE; Williams, 2001) for second through sixth grade range from .64 in sixth grade to .77 in fourth grade (Good et al., 2013).

A number of studies have demonstrated the validity of R-CBM as a screening measure, particularly at the elementary grades. In a study conducted by Shapiro, Keller, Lutz, Santoro, and Hintze (2006), researchers evaluated convergent validity of R-CBM measures with standardized state tests employed in Pennsylvania, as well as several standardized, norm-referenced tests. These included the SAT-9 (Harcourt Brace Educational Measurement, 1996), MAT-8 (Harcourt Brace Educational Measurement, 2000), and the Stanford Diagnostic Reading Test (SDRT; Karlsen & Gardner, 1995). A total of 1,048 students in grades three through five enrolled in two school districts were included in the evaluation of R-CBM

For fall, winter, and spring screening assessments, correlations with the state test ranged from .62 to .69 for all but the fall screening assessment in one of the two districts. In terms of

the relationship to published, norm-referenced assessments of reading, correlations with R-CBM screening assessments ranged from .70 to .72 for the MAT-8 Total Reading composite, and from .62 to .74 for the SAT-9 Total Reading composite.

A study examining predictive validity of a different set of published R-CBM measures found strong support for these measures as well (Nese, Park, Alanzo, & Tindal, 2011). Researchers examined the ability of easyCBM passages (Alonzo, Tindal, Ulmer, & Glasgow, 2006) to predict performance on the Oregon state achievement test for approximately 3,600 students in forth and fifth grades. Students were administered R-CBM passages during the spring benchmark, and then completed the Oregon state achievement test that same spring. Results indicated R^2 values of .71 for predicting fourth graders' state testing performance, and .70 for predicting fifth graders' performance. Numerous other studies have established that standardized measures of oral reading demonstrate strong relationships with both state and published standardized assessments of reading achievement, particularly in grades three through five (e.g., Keller-Margulis, Shapiro, & Hintze, 2008; Pearce & Gayle, 2008; Reschly, Busch, Betts, Deno, & Long, 2009).

Over the past few years, there has been an emphasis on research that examines the diagnostic accuracy of R-CBM. Studies reviewed indicate that the field is currently investing a great deal of effort into evaluating the predictive accuracy of R-CBM measures. This comes on the heels of a study by Jenkins, Hudson, and Johnson (2007), who reviewed available literature and found unacceptably low rates of sensitivity (correctly identifying those at risk) and specificity (correctly identifying those not at risk). The authors presented recommendations for sensitivity rates of 90-95% in predicting risk status below the 25-30th percentile on some outcome measure with acceptably high rates of specificity (Jenkins, Hudson, & Johnson, 2007).

Test developers are now considering predictive accuracy, which is evident in examinations of technical manuals for the DIBELS Next and AIMSweb measures. The AIMSweb manual presents results of Receiver Operating Characteristic (ROC) curves analysis for groups of students in grades 3 through 8 (Pearson, 2012a). Area under the curve values ranged from .83 to .94 (1.0 represents perfect classification). Sensitivity ranged from .77 to .80, while specificity ranged from .73 to .91 (Pearson, 2012a). The DIBELS Next technical manual (Good et al., 2013) indicated that when using the "at or above benchmark" criterion, average sensitivity and specificity rates were .49 and .99 respectively when predicting an intensive need for support at the end of the year. Using the "well below benchmark" criterion, average rates were .74 and .94 when predicting an intensive need for support at the end of the year (Good et al., 2013).

A series of independently conducted studies evaluating the predictive accuracy of various R-CBM screening measures suggests that these instruments are demonstrating improved rates of accuracy in predicting future risk status, thereby increasing the validity of R-CBM measures. Using a large sample of third graders from Florida's *Reading First* schools ($N = 35{,}207$), Roehrig, Petscher, Nettles, Hudson, and Torgesen (2008) compared DORF screening results from three different time points (September, December, February/March) to results of the Florida state reading assessment and a standardized norm-referenced reading achievement test. The total sample was also divided into two statistically equivalent groups, with one serving as the calibration sample (S1) and the other serving as the cross-validation sample (S2) for developing alternative cut scores.

Practical effects of the DORF measures were examined using receiver operating characteristic (ROC) curve analyses. Overall classification accuracy in predicting success in

meeting the end-of-year benchmark on the state test was .81 (September), .82 (December), and .83 (Feb/March). Sensitivity was more variable using original scores, with 74% of all those who failed to meet proficiency correctly identified as at risk in September, 88% in December, and 91% in Feb/March. Specificity at each point was .86 (September), .80 (December), and .81 (Feb/March).

The authors found that recalibrated cut scores based on ROC curves analysis were more effective in identifying students at risk. For the fall DORF assessment, sensitivity improved to .83, while sensitivity rates using cut scores from December and Feb/March assessments were slightly higher (.93 and .94 respectively). Using the cross-validation sample (S2) to examine cut scores indicated similar results (sensitivity of .84, .92, and .94 for September, December, and Feb/March assessments respectively). Specificity for recalibrated cut scores ranged from .75 to .87 at each assessment, with very similar rates for the cross-validation sample. Overall classification accuracy improved slightly when using recalibrated cut scores versus original cut scores, increasing from .78 to .86.

Similarly, Goffreda, DiPerna, and Pedersen (2009) used logistic regression analyses to determine predictive accuracy of first graders' ($N = 67$) winter DIBELS scores when outcome measures were a standardized norm-referenced test at the end of second grade, and the state test at the end of third grade. ROC curves analysis was used to determine optimal cut points. Using recommended cut scores sensitivity was .80, while specificity was .87 in predicting future performance on the standardized norm-referenced measure of reading ability. These were also the rates using optimal cut scores based on ROC curves analysis. When predicting proficiency on the state reading assessment, DORF sensitivity was .77, while specificity was .88. ROC

curves analysis yielded cut scores that produced higher sensitivity (.88) rates, and the same rate of specificity (.88).

A series of other studies using ROC curves analysis have found similar results. For instance, Petscher, Kim, and Foorman (2011) found that DORF demonstrated sensitivity rates of 60% and 66% when predicting risk status (performance below the 25th percentile) on two different norm-referenced measures of reading achievement. Specificity for these measures ranged from 81% to 87%. In another study, DORF screening measures in the fall had high rates of sensitivity (.88-.97) when predicting proficiency on the Pennsylvania System of School Assessment (PSSA) in the spring (Shapiro, Solari, & Petscher, 2008). Specificity ranged from .49 to .58. Finally, in a study on AIMSweb R-CBM measures, sensitivity was .71 when predicting fifth grade PSSA scores from fourth grade spring screening assessments, while specificity was .78 (Keller-Margulis, Shapiro, & Hintze, 2008). Predicting performance on state tests two years in advance for grades 1, 2, and 3 yielded sensitivity rates ranging from .72 to .79, while specificity ranged from .81 to .90.

The number of studies investigating predictive accuracy of R-CBM screening measures within the past few years is a clear indication of its importance to the field. Clearly, researchers are working to improve these measures in order for results to be useful in educational decision-making. Measures of oral reading appear to be generally efficient in correctly identifying students who are and are not at risk. Unfortunately, many studies are still not demonstrating the level of sensitivity (90-95%) in detecting risk that has been recommended by Jenkins, Hudson, and Johnson (2007). The amount of research being conducted in this area is a huge benefit to R-CBM screening measures, which require research advances in order to improve their instructional utility.

Scientific Issues with Current Curriculum-Based Measures for Screening Early Readers

Unfortunately, research is suggesting that current universal screening measures for reading in the early elementary grades demonstrate a host of scientific issues. On the one hand, current measures of oral reading do not maintain the psychometric strength at the early elementary grades that they do with students in the upper elementary grades. At the same time, psychometric limitations of many alternative early literacy universal screening measures make it difficult to justify their use as well.

Scientific Issues with Reading CBM

One clear example of a scientific issue with R-CBM at the early grades is found in results of research conducted by Catts, Petscher, Schatschneider, Bridges, and Mendoza, (2009), which indicated substantial floor effects, and therefore poor predictive validity, for R-CBM screening results in first grade. Specifically, DIBELS ISF, LNF, PSF, NWF, and ORF measures were evaluated for a cohort of 18,667 students as they traveled from kindergarten through second grade.

Visual inspection of frequency histograms indicated that DORF screening measures, which were administered beginning in the fall of first grade, were characterized by very strong floor effects. The histograms were strongly positively skewed, with the majority of students scoring at the low end of the possible score range. Furthermore, while distributions tended to normalize, this did not happen until approximately the end of second grade. Overall, this indicates that strong floor effects characterize DORF screening assessments administered throughout first, and some of second grade, with few students having the skills necessary to perform in the middle or upper score ranges (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009).

Additional investigations indicated that these floor effects significantly impacted the predictive validity of the screening measures. Correlations between DORF screening measures and the outcome measure, DORF in third grade, were quite low, and tended to improve as the score distributions normalized (after several administrations; Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009). Similar results were found in a study by Petscher and Kim (2011), who observed the largest standard deviations for DORF in first grade, compared to second and third grade. These findings are important because they highlight problems with R-CBM for young students that impact validity. Considering the importance of early intervention for later learning outcomes (e.g., Snow, Burns, & Griffin, 1998; Torgesen, 1998), this is an essential issue to address in future research and measurement development.

Scientific Issues with Other Curriculum-Based Measures

Although a number of alternative measures are available for universal screening in kindergarten and first grade, many demonstrate serious psychometric issues that limit their utility. In the study described earlier conducted by Catts, Petscher, Schatschneider, Bridges, and Mendoza, (2009), researchers found that floor effects were present in most every early literacy screening measure. Of the DIBELS measures administered in kindergarten, ISF demonstrated strong positive skews at each administration until it was discontinued in February of kindergarten (third administration). The LNF histograms were also positively skewed in the beginning of kindergarten, but normalized by the spring (April) administration of kindergarten. Both the DIBELS PSF and NWF measures also demonstrated positively skewed histograms when they were first administered in the middle of kindergarten. However, frequency histograms remained positively skewed well into first grade for both measures. While PSF

demonstrated normalization by February of first grade, NWF continued to demonstrate strong floor effects well into second grade.

In terms of diagnostic accuracy, measures other than R-CBM in these early grades perform quite poorly. In a study conducted by Johnson, Jenkins, Petscher, and Catts (2009), the researchers found extremely disappointing results for ISF, PSF, and NWF. Participants included 12,055 students who were followed from kindergarten through third grade. DIBELS ISF, PSF, NWF and ORF screening measures were administered periodically across the four years. Criterion measures included the Florida state achievement test, the SAT-10, and the Peabody Picture Vocabulary Test – Third Edition (PPVT; Dunn & Dunn, 1997). When sensitivity was set to .90, classification accuracy was 41% for PSF in first grade, and 57% for NWF. When allowing sensitivity and specificity to be driven by improvements in classification accuracy, sensitivity plummeted to .2 and .38 for PSF and NWF respectively. Results were similar for kindergarten measures.

In a retrospective study to examine diagnostic accuracy Clemens, Hilt-Panahon, Shapiro, and Yoon (2012) evaluated the ability of DIBELS ISF, LNF, PSF, and NWF assessments in kindergarten and first grade to predict DORF performance at the end of first grade. Data for a group of 101 students was collected for kindergarten through first grade on screening measures, administered yearly in the fall, winter and spring. Although LNF and NWF scores demonstrated acceptable average AUC values (.82 for LNF, .845 for NWF), ISF and PSF did not (.70 and .65 respectively). Additionally, the authors found that there was a decrease in PSF AUC values across time, indicating a decline in diagnostic accuracy. When using slopes to predict DORF performance at the end of first grade, of the kindergarten measures only LNF produced adequate strength, while NWF was the only measure to demonstrate this predictive utility in first grade.

Furthermore, in a synthesis of the literature on early literacy screening measures, Goffreda and DiPerna (2010) found limited support for the use of available screening measures intended for kindergarten and first grade. Across the studies reviewed, the authors found limited concurrent and predictive validity evidence for ISF, PSF, and NWF. More promising predictive validity evidence exists for LNF. The authors also noted that existing research suggested that DORF scores are the most accurate of all DIBELS measures in predicting future reading difficulties, and that they tend to over identify students as at risk.

A major source of psychometric issues with current early literacy screening measures may be the fact that they represent such specific skills. As Fuchs (2004) points out, the purpose of CBM was to develop measures that would be psychometrically sound as a result of their emphasis on identifying indicators that require mastery of multiple skills. Issues such as floor and ceiling effects, for example, are more likely to occur with specific skills that are learned relatively quickly. Additionally, Fuchs notes that there is a lack of evidence to support the use of single skills as indicators of general curricular outcomes. As a result, measures such as LNF and NWF may offer little or no useful information about academic development, and may also lead teachers to narrow their instructional techniques to the disadvantage of their students.

Practical Issues with Current Curriculum-Based Measures for Screening Early Readers

Similar concerns have been voiced by a number of individuals who object to the use of existing early literacy screening measures, such as DIBELS. For instance, Pearson (2006) has concluded that these measures "shape instruction in counterproductive ways by directing schools and teachers to a limited set of features of the reading curriculum" (p. x). In the same vein, Goodman (2006) asserts that many tasks, such as letter naming and nonsense syllable reading, do not actually represent the big ideas of reading that they purport to measure, because they reduce

them to such small component parts. Additionally, he criticizes the authors of these measures for assuming that early literacy measures, including ISF, LNF, PSF, and NWF build successively upon one another, thereby providing a comprehensive picture of the road to reading proficiency. For example, he asks how it is that fast letter naming contributes to reading development over simply knowing the letters or not.

Goodman (2006) also raises the issue that dialect influences speech and a student's ability to hear and segment words in the phoneme segmentation fluency assessment. Furthermore, he notes that the NWF measure includes some real English and Spanish words, which presents the issue that children may pronounce some words in a specific dialect and be scored incorrectly. Finally, this assessment may unfairly punish students who read non-words as if they are the real words they resemble, and presents a host of words that violate English spelling rules, such as words that end with *j*.

The NWF measure, in particular, has raised concern with researchers and educators alike. In addition to Goodman's (2006) criticism, a survey and interview study revealed that educators view the use of nonsense words as a disadvantage of the DIBELS assessments (Hoffman, Jenkins, & Dunlap, 2009). Researchers mailed surveys assessing educators' opinions of the DIBELS measures to members of the state council of the International Reading Association and received 87 responses (24% response rate). Respondents included primarily classroom teachers (51%), as well as reading specialists, special educators, administrators, and university faculty. Seven individual interviews were also conducted with local school personnel familiar with the DIBELS.

Results of the survey responses indicated diversity in opinion about advantages and disadvantages of using the DIBELS (Hoffman, Jenkins, & Dunlap, 2009). Of the advantages

respondents noted for the DIBELS, 44% indicated that it is quick and easy to use, 21% indicated it identifies at-risk readers, and 20% noted that it informs instruction. At the same time, 17% of respondents indicated that administration time was a *disadvantage*, as a result of the individualized testing required. Some other disadvantages noted included feeling that the information yielded was restricted (17%), that comprehension was not tested (16%), and that nonsense words were used in assessment (5%). Individual interviews supported these results. In particular, interviewees repeatedly articulated concerns regarding the amount of time spent on individual repeated administrations of various DIBELS tests.

Acceptability information has offered useful information about assessments from the perspective of consumers. In the past, CBM has been rated more acceptable than standardized, norm-referenced assessments of academic skills. In one study, Eckert and Shapiro (1999) examined the acceptability of R-CBM data versus standardized cognitive and achievement data for a hypothetical student case. Raters included 631 general elementary education teachers of first through fifth grade from across the country. Of the participants, 418 completed either a rating of their acceptability of one case or the other (CBM data or standardized assessment data), while 201 participants completed acceptability ratings of both hypothetical case descriptions. Regardless of condition (between-subjects or within-subjects), CBM data were found to be significantly more acceptable to teachers than standardized assessment data.

Information indicating that R-CBM data is preferable to teachers over standardized assessment data must be reconciled with the fact that persistent concerns exist regarding CBM at the early elementary grades. Specifically, although R-CBM data appears to be generally acceptable to teachers (Eckert & Shapiro, 1999), these measures are psychometrically limited or unavailable in kindergarten and first grade. Additionally, it appears that the DIBELS measures

that do exist for these students continue to raise concerns and attract criticism (e.g., Goodman, 2006; Hoffman, Jenkins, & Dunlap, 2009). Furthermore, a limitation of the current literature of acceptability of CBM measures is the lack of comparisons between different forms of CBM measures, rather than between CBM and standardized tests. To the author's knowledge, no studies have directly compared teacher acceptability ratings of different early literacy CBM screening measures.

Possible Alternatives to Current Curriculum-Based Measures for Screening Early Readers

Recognizing limitations of current early literacy CBM screening measures, researchers have begun to investigate alternative CBM approaches to assessing reading ability in kindergarten and first grade students.

Word Identification Fluency

One approach is Word Identification Fluency (WIF), where students are asked to read from a list of words presented in isolation from connected text for one minute. This approach was included as one of the procedures investigated by Deno, Mirkin, and Chiang (1982), and has been the focus of more recent research in light of concerns over existing screening measures for younger students.

Direct comparisons of WIF and other early literacy CBM screening measures has shown that WIF is superior in its ability to identify students at risk of later reading failure. In an examination of concurrent and predictive validity, results favored WIF over NWF for nearly all analyses (Fuchs, Fuchs, & Compton, 2004). Researchers assessed 151 first grade students from 33 classrooms considered at risk for reading difficulties based on performance on a letter naming fluency task. Participating students were assessed in the fall and spring using the WIF and NWF measures. In addition, progress monitoring data was collected using both measures weekly for 7

weeks and twice weekly for 13 weeks. Standardized assessments of word identification, nonsense word decoding, reading fluency and comprehension served as criterion measures. Correlations indicated WIF maintained superior concurrent and predictive validity when considering measures of word identification, reading fluency, and reading comprehension as criterion measures. Both WIF and NWF demonstrated moderate correlations with fall and spring measures of nonsense word reading. Dominance analyses, an extension of multiple regression, also indicated that WIF was superior to NWF for predicting reading fluency and reading comprehension at the end of first grade ($p < 0.05$).

Using ROC curves analysis, WIF has also demonstrated superior diagnostic accuracy when compared to LNF, PSF, and NWF (Clemens, Shapiro, & Thoemmes, 2011). In a study of 138 first graders, participating students were assessed in the fall on DIBELS LNF, PSF, NWF, and WIF. In the spring, students were evaluated again on a series of outcome measures, including DORF, the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999) Sight Word Efficiency and Phonemic Decoding Efficiency subtests, and Maze. Results indicated that WIF was a significant predictor of all outcome measures, and that none of the other screening measures contributed significantly to predictive accuracy when TOWRE subtests or Maze served as the outcome variables. Only the PSF measure contributed significantly ($p < 0.01$) when predicting DORF.

When sensitivity was set to approximately .90 for predicting reading difficulties (performance below the 30th percentile on outcome measure) WIF consistently demonstrated the highest AUC value compared to other individual screening measures (LNF, PSF, NWF; Clemens, Shapiro, & Thoemmes, 2011). Values ranged from .862 when predicting spring performance on the TOWRE to .909 when predicting performance on a latent variable composite of all three

outcome measures. The WIF measure also demonstrated relatively high levels of sensitivity (.52 to .71). Only LNF demonstrated a comparable range (.56-.69). Combining WIF with other screening measures resulted in only modest improvements to classification accuracy in general.

In another study of first grade students, Speece and colleagues (2011) found similarly promising results for measures of word reading. Researchers examined the ability of single scores in the fall and slopes for 243 first graders to predict end of year reading performance on a variety of criterion measures. The model that best predicted end of year reading achievement included the fall TOWRE Sight Word Efficiency score, WIF, and a teacher rating of reading problems. ROC curves analysis yielded an AUC value of .96 for the combined ability of these three measures to predict reading outcomes at the end of the year.

Decodable Text

Clearly, research is demonstrating that WIF is superior to commonly employed early literacy CBM screening measures, such as LNF, PSF, and NWF. However, considering criticism from individuals such as Pearson (2006) and Goodman (2006) regarding the reductionist approach to reading that these measures take, it is unlikely that WIF would provide an acceptable alternative. Furthermore, studies have demonstrated superiority of R-CBM assessments of word identification (e.g., Deno, Mirkin, & Chiang, 1982). Additionally, models of reading development indicate that first grade represents a time of substantial growth in decoding skills and sight word vocabulary that is applied to reading connected text (Ehri & McCormick, 1998; Spear-Swerling & Sternberg, 1994). Finally, if the goal is to develop proficiency reading connected text, it is logical to assess that skill over less directly related skills.

A solution may be to develop passages that include decodable text and developmentally appropriate sight words. According to Mesmer (2001), decodable text is defined by the presence

of two features. First, the text includes a high proportion of phonically regular words, or words that have phonically regular relationships between the component letters and their sounds. Second, decodable text includes a match between the letter-sound relationships included in the text, and those the student has learned. Considering that R-CBM in first grade has demonstrated substantial floor effects (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009), passages developed around the phonics skills and sight word vocabularies that first graders have been taught may offer a successful CBM approach to universal screening at this age. In particular, if current passages include words that are simply too difficult for most students to learn, a way to capitalize on the psychometric advantages of R-CBM for older elementary students is to make them developmentally appropriate for younger students.

Studies have indicated that such an approach may be superior to WIF for universal screening. For example, in a study with fifth grade students Klauda and Guthrie (2008) assessed 278 students using external criterion measures of reading comprehension [Gates-MacGinitie Reading Test (GMRT; MacGinitie, MacGinitie, & Dreyer, 2000), reading fluency [Woodcock-Johnson III Reading Fluency subtest (WJ-III; Schrank, Mather, & Woodcock, 2004), and researcher-developed WIF and R-CBM], background knowledge, and inferencing in the fall and winter. Twelve weeks later, participating students completed the GMRT Comprehension subtest, the WJ-III Reading Fluency subtest. Results indicated that various reading fluency skills assessed contributed uniquely to reading comprehension. Specifically, fluent recognition of individual words, fluency reading connected text (both silently and aloud), and expressive oral reading ability all predicted student comprehension.

In a study of first grade students, Lopez, Thompson, and Walker-Dalhouse (2011) found that proficient readers demonstrate greater fluency with connected text than do average and less

skilled readers. Participants included 283 students who were assessed at three points across the year (fall, winter, spring) using measures of WIF and researcher developed R-CBM passages that included the same words as the WIF measures. Students were designated as either less skilled, average, or proficient readers based on end of year R-CBM scores.

Results indicated an interaction between skill level and reading fluency on words in context versus isolation (Lopez, Thompson, & Walker-Dalhouse, 2011). Specifically, skilled readers appeared to rely on the context of connected text to aid in their reading. These students scored significantly higher when words were presented in a passage format than when reading the words in isolation. Average readers read words faster in isolation at the beginning of the year, at about the same rate as words in context in the middle of the year, and more slowly than connected text by the end of the year. Less skilled readers demonstrated more fluent reading when words are presented in isolation than in context at each assessment. The authors concluded that this indicated proficient readers use the context of connected text to aid in their reading fluency.

Decodable text may offer further advantages for assessing all learners because it presents words that are connected to the specific curricular goals of first grade students. In her study on decodable text, Mesmer (2005) examined performance on decodable versus less decodable texts for a group of 23 first grade students receiving a two week phonics intervention. Following each 20-minute daily lesson, students read either highly decodable or less decodable text. Results indicated that students reading highly decodable text demonstrated more frequent application of correct phonics skills ($p < 0.05$), read more accurately ($p < 0.05$), and relied on the examiner to supply words less frequently ($p < 0.05$).

Together, these data suggest the potential for decodable text to be a successful CBM reading screening measure for first grade students. While theories and studies of reading development indicate the importance of decoding skills and sight word knowledge at this level, studies indicate that assessments of these skills are predictive of later reading ability. A measure that can combine the assessment of these skills (phonics and sight word knowledge) in a way that also evaluates oral reading with connected text should offer a significant advantage over current measures.

Chapter III: Method

Participants

Participants in the study included 234 first-grade students in 15 classrooms from 4 elementary schools serving students in kindergarten through fifth grade in 2 school districts in Eastern Pennsylvania. The author and her advisor contacted district administrators (e.g., superintendents, assistant superintendents) in districts with which they are familiar. Initial contact was made via email and phone, with on-site visits as necessary. Only those schools that were implementing universal screening using CBM with first-grade students were eligible to participate.

Following district and building-level approval, all first-grade teachers in the participating elementary schools were provided with information about the study, and given the opportunity to ask questions. Demographic information including gender, age, race/ethnicity, and years of teaching experience was collected for participating teachers in each district using an electronic survey completed by each teacher. The researcher also documented the reading curriculum being employed with first-graders at each school. Three of the four participating schools were utilizing the Treasures curriculum published by McMillon-McGraw Hill, while the fourth school utilized the Reading Street curriculum offered through Scott Foresman. The final sample of teachers included 15 first grade educators. Of these teachers, 93% were female (7% male), 93% were white, and 7% were mixed race/multiracial. The average age of participating teachers was 43 years, ranging from 26 to 66, and average years of teaching experience was 15, ranging from 2 to 34.

A letter requesting parental consent for student participation was sent home for all students in the targeted first-grade classrooms. The letter included a short description of the study, researcher contact information, and information pertaining to participant rights, as well as

potential risks and benefits. Students were eligible to participate following receipt of written parental consent and verbal student assent.

Students routinely excluded from the schools' screening process (e.g., students with significant disabilities who are unable to participate in the assessment) were excluded from the study. Student demographic information was also collected for participating students including gender, age, race/ethnicity, free/reduced lunch status, English language learner (ELL) status, provision of special education services. Permission to obtain this information from each child's school was included in the consent form sent home to parents. Small prizes (e.g., pencils, stickers) were offered to all students who return a completed consent form, regardless of the families' choice to allow their children to participate or not.

The final student sample included 234 first grade students ranging in age from 6 years, 2 months to 8 years, 2 months, with an average age of 6 years, 9 months. Just over half (58.55%) of students received free or reduced priced lunch, and approximately half (50.85%) of the students were male. The sample also included a small number of students receiving special education services (5.98%), while another small group (8.55%) participated in English as a Second Language (ESL) classes. The final student sample was racially/ethnically diverse, including 45.30% White/Nonhispanic, 35.04% Hispanic, 7.26% Black/African/Jamaican/West Indian, 5.98% multiracial, 4.70% Asian, 0.85% American Indian, and 0.43% Native Hawaiian/Pacific Islander. Attrition was minimal over the course of the study. By the spring assessment, 12 students had left the study due to moving, resulting in their being withdrawn from their participating elementary school. These students were not assessed during the spring HD passage assessment. Two additional students moved away before GRADE assessments took place in the spring of 2014. Therefore, complete data (fall HD passage assessments, spring HD

passage assessments, and GRADE assessments) were collected with a total of 220 first grade students. DIBELS assessment results, which were supplied by school administrators, were available for a total of 216 students in the winter and for 215 students in the spring.

A group of 20 classroom teachers were also recruited to participate in the acceptability portion of the study. Teachers in each of the participating first-grade classrooms were provided with the option of participating in a short electronic survey. Additional teachers at elementary schools in Pennsylvania and New York State were also contacted and offered the opportunity to complete the electronic survey. Consent was obtained electronically for all participating teachers at the start of the survey, as well as demographic data, including age, gender, race/ethnicity, and years of teaching experience. The sample of teachers ranged in age from 23 to 57 years (M = 42.0 years). All teachers were female, and 95% were White, while 5% were Black/African American. The years of teaching experience ranged from 1 to 31 with a mean of 14.3 years. Of that time, teachers reported instructing first graders for an average of 7.8 years (range = 1–25). A total of 70% of the teachers in the sample resided in Pennsylvania, while 30% resided in New York.

Setting

All individual student assessments took place in a separate room or other area, away from other students, to avoid distractions during testing. Group assessments were conducted either in the students' classrooms, or in another quiet area, such as an empty cafeteria or library. Students not participating in the study were removed to other areas for separate instruction while testing took place. During group assessments, students sat at individual desks or large tables, and efforts were made to reduce all noise and other distractions. Teachers who completed the acceptability

rating scale did so on a computer at their convenience within 2 weeks of receiving the electronic survey link.

Measures

Highly Decodable Passages. Highly decodable passages (HD passages; Shinn, 2009; 2012) are brief passages of approximately 200 words that include a high proportion of wholly decodable words and developmentally appropriate sight words. Wholly decodable words are those that typically include only single consonants and common vowel sounds (commonly short vowel sounds) such as the words /cat/ and /big/. The process used to develop the HD passages began by having a former kindergarten and first-grade classroom teacher develop an initial draft of passages, which were then submitted to expert review by researchers and word analysis software. This process ensured that the final set of 10 passages were unique, logical, and included a high proportion (approximately 75%) of highly decodable words, as well as a smaller number of sight words and non-decodable words (see Appendix A).

The procedure used to administer HD passages involves having assessors follow R-CBM procedures. The student is asked to read aloud while being timed for 1 minute. As the student reads aloud, the assessor marks any words read incorrectly. Words are considered incorrect if a student mispronounces the word, skips the word, replaces the word with a different word, or hesitates for 3 seconds or more. At the end of 1 minute the total number of words read correctly is considered the student's score on that passage. A total of 3 passages are administered, with the median selected to represent the student's oral reading fluency score in order to account for passage variability.

Initial field testing indicated that HD passages have strong reliability and validity for use with first graders (Shinn, 2009; 2012). Alternate form reliability ranged from .91 to .96, with a

median paired score correlation of .93. Initial concurrent validity was investigated with oral reading fluency passages (R-CBM) and nonsense word fluency (NWF) at the middle of the year first grade screening assessment. Correlations were strong for both R-CBM (.88) and NWF (.71) (Shinn, 2009; 2012). For the current study, the author and her advisor consulted with the developing researcher to identify the 3 passages administered from the full set of 10 existing HD passages. These passages demonstrated reasonably equivalent means and strong reliability based on previous examinations with kindergarten and first grade students.

Dynamic Indicators of Basic Early Literacy Skills. The Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2011) is a set of screening measures for students in kindergarten through sixth grade. The measures are intended to be used to identify students at risk for future reading difficulties, assist teachers in identifying areas in which to provide instructional support, monitor students' progress toward reading goals, and examine school-wide instruction and intervention effectiveness (Good et al., 2013).

Nonsense Word Fluency. Nonsense Word Fluency (NWF) is a 1-minute, timed assessment of the alphabetic principal and early phonics skills. During the assessment, students are asked to read aloud from a list of phonetically regular nonsense (not real) words that follow a consonant-vowel-consonant or vowel-consonant pattern. The assessor underlines any correctly read letter sounds, either isolated letter sounds or sounds that are blended together. Two separate scores are calculated for NWF, Correct Letter Sounds (CLS), or sounds any sounds read correctly in isolation or blended together, and Whole Words Read (WWR), or the number of nonsense words read correctly in their entirety. First-grade alternate form reliability (single form) has been reported as .85 for CLS and .90 for WWR; test-retest reliability is .76 for CLS and .70 for WWR. Inter-rater reliability was reported as being strong in first grade (.99 for both

CLS and WWR). Predictive validity coefficients when predicting end of year performance on the GRADE were found to be moderate in the fall (.43 for CLS, .39 for WWR) and stronger in the winter (.51 for CLS, .52 for WWR; Good et al., 2013).

DIBELS Oral Reading Fluency. The DIBELS Oral Reading Fluency (DORF) measure is used to assess students' speed and accuracy reading connected text. Students are asked to read three grade-level passages for 1 minute each. While the student reads, the assessor marks errors (substitutions, omissions, hesitations of more than 3 seconds, and incorrectly read words). The median words read correct and median errors are taken to represent the student's performance. Accuracy can also be calculated using these scores [median words correct/(median words correct + errors)]. Alternate form reliability for words read correct using single forms is reported to be .95 at first grade. Using three DORF passages, alternate form reliability is reported to be between .97 and .98 for words read correct at first grade. Two-week test-retest reliability for DORF in first grade is .95. Predictive validity with end of year performance on the GRADE was reported to be .64 for winter DORF assessments, while concurrent validity was reported to be .75 (Good et al., 2013).

Group Reading Assessment and Diagnostic Evaluation. The Group Reading Assessment and Diagnostic Evaluation (GRADE; Williams, 2001) is an untimed, group-administered, norm-referenced assessment of reading abilities for students in grades prekindergarten through early postsecondary level; in first grade, a series of tests address comprehension and word knowledge. Reliability for the GRADE is strong, with internal consistency ranging from .89 to .99 and alternate form reliability ranging from .81 to .94. Test retest reliability ranged from .77 to .98. Concurrent validity with the California Achievement Test has also been reported to be strong, ranging from .82 to .87 for levels 1 and 2 (first and

second grade). Predictive validity with the TerraNova (administered in the spring) ranged from .76 to .86 using fall GRADE performance for levels 2, 4, and 6 (Waterman, 2012).

Assessment Rating Profile-Revised. The Assessment Rating Profile-Revised (ARP-R; Eckert, Hintze, & Shapiro, 1999) is a brief 12-item rating scale used to evaluate the acceptability of assessment instruments. Raters are asked to indicate the degree to which they agree or disagree with a given statement on a Likert scale from 1 (strongly disagree) to 6 (strongly agree). Statements such as "I liked the assessment procedures used in this assessment," and "overall, this assessment would be beneficial for the child" are included. Internal consistency is strong (.99), as well as test-retest reliability (.82 to .85; Eckert, Hintze, & Shapiro, 1999).

Procedures

The entire student sample (234 students in the winter and 220 students in the spring) were assessed at two points during the 2013-2014 school year, once in the winter (December) and once in the spring (April), in order to be commensurate with screening schedules established in the schools. As noted previously, all participating schools were conducting universal screening three times per year (fall, winter, spring) using the DIBELS Next. Therefore, participating first graders were administered typical winter and spring screening measures by school staff with DIBELS Next NWF and DORF in the winter, and again in the spring. Graduate students in school psychology were trained to administer HD passages. Training included a brief description of HD passages and their development, followed by training on steps for administering and scoring HD passages. Finally, graduate students practiced scoring passages using audio recordings, after which they compared their ratings to a scoring key. Finally, inter-rater agreement was collected using a final audio recording to ensure all data collectors demonstrated at least 90% agreement. Three passages were administered to students in the

current study in the winter, and then administered again in the spring. At both time points, the median score was selected to represent the student's oral reading performance, and scores on all three passages were retained in order to examine alternate form reliability on the three HD passages administered.

Approximately 4 weeks after the winter assessment, a subsample of 100 first grade students, randomly selected from the entire sample, were selected to be administered the three HD passages a second time to assess test-retest reliability. Although students were originally planned to be re-tested 2 weeks after the initial HD passage assessment, a 4-week gap was necessary as a result of winter recess, which occurred for two weeks at the end of December and beginning of January. Due to absences, a total of 96 students were re-tested.

Ten percent of all DIBELS Next assessments and HD passages were randomly selected for inter-rater agreement calculations. In order to do this, assessments were audio recorded in order for the researcher to independently score each assessment for the selected students. Word-by-word agreement [agreements/(agreements + disagreements)] was calculated for each passage and the average agreement was then computed. Results of inter-rater agreement calculations indicated highly reliable DIBELS Next data collection. Overall, inter-rater agreement was 95.8% across all DIBELS Next measures (NWF CLS, NWF WWR, and DORF). Average agreement for NWF CLS was 95.4%, while NWF WWR was calculated at 89.7%. Average DORF inter-rater agreement was calculated to be 98.0%. Results of inter-rater agreement calculations for HD passages are presented in the results section of this document, and also indicated highly reliable data collection.

During the spring assessment period, all participating first graders were also administered the GRADE Comprehension Composite. This testing was conducted by the lead researcher, as

well as trained graduate assistants. Students participated in this group-administered assessment in their classrooms or another designated testing area, and received stickers for completion of the assessment.

To examine the acceptability of the HD passages compared to DIBELS Next NWF, first-grade teachers in participating schools were asked to complete brief acceptability surveys. Teachers who consented to participate received an email with the link to the electronic survey. Additionally, teachers from various elementary schools in New York and Pennsylvania received the email with the survey link and a request to participate following administrative approval. Participating teachers were asked to read two short descriptions of the assessments and resulting data for a hypothetical first grade student. They were then asked to complete the 12 items of the ARP-R based on each assessment description. Surveys were counterbalanced across participants, so that half of all participants rated the HD passages first and DIBELS Next NWF second. A second group of participants rated the measures in the reverse order. This survey was completed at each teacher's convenience within a two-week window.

Data Analyses

The purpose of the current study is to examine the psychometric properties of a newly developed measure, HD passages. As such, reliability, convergent validity, and predictive accuracy were investigated. A secondary research question examining teacher acceptability of the measures was also examined. The following steps were taken to analyze the research questions.

A priori analyses. Analyses were conducted a priori to determine sample sizes necessary for conducting each of the analyses included. The minimum sample size identified for Pearson correlations conducted as part of this study was 100 based on research that indicates bias occurs

in smaller samples (Wang & Thompson, 2007). According to de Leeuw and Kreft (1995), a sample should include at least 20 groups (in this case classrooms) with at least 5 group members (students) in order to complete hierarchical linear modeling (HLM). Based on research by Peduzzi and colleagues (1996), a sample of 167 first grade participants was identified as the minimum necessary for conducting logistic regression, and the subsequent ROC curves analysis to examine predictive accuracy. With an anticipated effect size of 0.5 (medium effect size), power set to 0.80, and a *p*-value of 0.05 it was determined that a sample size of 34 was necessary for a dependent means *t*-test.

In terms of Pearson correlations, the minimum sample size was achieved for all but test-retest correlations. However, given that the final sample for this set of analyses was very close to the goal of 100 (96 students), it was considered appropriate to continue with analyses as results should not be strongly affected. The sample size for HLM did not meet the recommended minimum outlined by de Leeuw and Kreft (1995), as the final number of classrooms was 15, rather than the recommended 20 (number of groups). Nor was the minimum sample size achieved for the dependent means *t*-test needed to analyze results of the teacher acceptability study. Therefore, sample size may have affected the ability to detect significant results with these analyses. The required sample size was exceeded for analyses required for investigating predictive accuracy, including logistic regression and ROC curves analysis, with a sample size of 220 students.

Preliminary analyses. Prior to running any statistical analyses, descriptive statistics, frequency tables, and histograms were examined to determine the extent and pattern to missing data, score ranges, and to check normality of the data. Skewness and kurtosis statistics were examined to ensure normality. Normal probability plots (P-P plots) were examined for

univariate normality, as indicated by a relatively straight line. Following recommendations by Stevens (2009), scatterplots were visually inspected to ensure bivariate normality, as indicated by an elliptical shape.

Analysis of research questions.

Q1: What is the reliability (i.e., inter-rater, test-retest, and alternate form reliability) of HD passages? Inter-rater reliability was calculated for 10% of all HD passages (randomly selected) using word-by-word agreement [agreements/(agreements + disagreements)]. Pearson Product Moment correlations were used to examine test-retest reliability for a subset of students (96) who were administered HD passages within approximately 4 weeks of the first assessment period (winter assessment). Pearson Product Moment correlations were also used to examine alternate form reliability for the three passages administered at the same point in time by correlating scores from the three passages administered (HD passage 1 vs. HD passage 2 vs. HD passage 3 at the winter, retest, and then spring assessments). Following recommendations outlined by Evans (1996), correlations were classified as strong ($\geq .70$), moderate (.40 to .69), or weak ($\leq .39$).

Q2: What is the convergent validity (i.e., concurrent and predictive validity) of highly decodable (HD) passages for assessing reading ability in 1st grade with currently available measures of reading ability (e.g., DIBELS Next NWF, DORF, GRADE)? Convergent validity, including concurrent and predictive validity, was calculated using Pearson Product Moment correlations. Specifically, to examine concurrent validity, Pearson Product Moment correlations were conducted between the HD passages score (median words read correct) and scores from various DIBELS Next screening measures at the same time point (winter and spring). Concurrent validity was examined using the same analysis for the GRADE in the spring only.

Using Pearson Product Moment correlations, predictive validity was examined by correlating HD passage scores from the winter administration with DIBELS Next screening measures and the GRADE (administered in the spring). Following recommendations outlined by Evans (1996), correlations were classified as strong ($\geq .70$), moderate (.40 to .69), or weak ($\leq .39$).

Q3: How well do students' fall HD passage scores predict their winter HD passage scores, and how does inter-classroom variability affect this relationship between winter and spring scores? Hierarchical linear modeling (HLM) was employed using HLM 7 (Raudenbush, Bryk, & Congdon, 2011) to investigate the relation between students' winter HD passage scores and their spring HD passage scores. This method was used to provide an understanding of how performance on these passages in the winter, as measured in words read correctly per minute, is associated with performance in the spring, and the degree to which classroom variability affects this relationship.

The advantage of using HLM is that it allows researchers to account for the nested data structure. In this case, 220 students are nested within 15 classrooms, where students share a similar environment and are likely to demonstrate similar scores on the HD passages as a result (Raudenbush & Bryk, 2002). Whereas the violation of the assumption that all observations are independent precludes the use of ordinary multiple regression, HLM allows for error terms to be correlated for nested groups (Hox, 2010).

Before the full model was run, an unconditional means model (i.e., excluding the winter HD passage score and classroom as predictors) was run to compute the intraclass correlation (ICC). When the ICC is substantial, it is inappropriate to ignore the classroom variability by running OLS regression. Therefore, HLM should be implemented.

Two levels were included in the model, including an individual student level (level 1), and a classroom level (level 2).

The full HLM will be mathematically given as:

Level 1 (student level; within classroom):

$$y_{ij} = \beta_{0j} + \beta_{1j} WinterHDPScore_{ij} + r_{ij} ,$$

Level 2 (classroom level):

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Classroom_j + \upsilon_{0j} ,$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Classroom_j + \upsilon_{1j} .$$

Combined, the 2-level HLM is given by:

$$y_{ij} = \gamma_{00} + \gamma_{01}\, Classroom_j + \upsilon_{0j} + (\gamma_{10} + \gamma_{11} Classroom_j + \upsilon_{1j})\, WinterHDPScore_{ij} + r_{ij}$$

At level 1, the intercept and slope for each individual student was estimated to determine if the HD passage scores from the winter assessment are significantly associated with the HD passage scores at the spring assessment (level-1 outcome). At level 2, the effect of classroom on the intercept (average spring HD passage scores, controlling for the effect of winter scores and the classroom variability) and the slope (growth from winter to spring) was assessed. The level-2 random effects are the departure from the average intercept and departure from the average pretest slope due to teacher/classroom variability, as well as the level-1 random error for each student in each classroom. The significance of the fixed and random effects were determined by a *p*-value of .05 or less.

Q4: What is the diagnostic accuracy (i.e., sensitivity, specificity, negative and positive predictive power, negative and positive likelihood ratios, post-test probabilities, area under the curve) of HD passages for 1st grade students? Using cross tabulations, sensitivity, specificity, positive predictive power, and negative predictive power were determined. Specifically, these

statistics were used to examine how well performance on the HD passages is able to predict students who go on to have reading difficulties, as defined by performance below the 40th percentile on the GRADE. There is a lack of agreement as to what the criterion for predicting "at risk" status should be, with researchers adopting different criteria in their research. For example, some researchers present performance below the 25th percentile as a possible criterion for determining unsatisfactory reading performance (e.g., Jenkins, Hudson, & Johnson, 2007). Others tend to set the bar a bit higher, indicating performance below the 40th percentile may be more appropriate (e.g., American Institutes for Research, 2007; Petscher, Kim, & Foorman, 2011).

Therefore, the decision was made to investigate diagnostic accuracy of HD passages when predicting to two different criteria for determining students who are "at risk" for future reading difficulties and to compare the results of these analyses. The two criteria included performance below the 25th percentile and performance below the 40th percentile. These two criteria allowed for examination of whether HD passages perform differently when predicting to different outcomes, and also to consider the use of HD passages for the specific needs of educators, with some settings opting to identify a higher number of students in need of intervention, while others may desire to identify only those with the greatest need. Using chi-square analyses, significance of these statistics was determined as having a *p*-value of .05 or less.

Sensitivity is defined as the degree to which a measure accurately identifies those students who go on to have difficulties in reading based on a future criterion measure (i.e., true positives), while specificity refers to the degree to which a screening measure accurately identifies those students who *will not* go on to have reading difficulties (i.e., true negatives; Jenkins, Hudson, & Johnson, 2007). Positive predictive power indicates the proportion of true

positives of all those identified by the screening measures as being at risk, while negative predictive power represents the proportion of students that the measure designated as not at risk who truly went on to have no reading difficulties (Furr & Bacharach, 2008).

Likelihood ratios and post-test probabilities were calculated following steps outlined by VanDerHeyden (2010). Likelihood ratios indicate whether or not the predictions made using the screening measure exceeded chance prediction, while post-test probabilities are computed from likelihood ratios and provide a single number that indicates whether the measure can improve diagnostic or predictive accuracy.

Finally, receiving operating characteristic (ROC) curves analysis was utilized to further examine the diagnostic validity of HD passages for predicting performance on the GRADE. By plotting the true positive rate (Y axis) against the false positive rate (X axis) we can generate a graphic representation of diagnostic accuracy. Because a straight line indicates identification at a chance rate, the area under the curve (AUC) serves as a valuable metric. AUC values, therefore, range from .5 (equivalent to chance) to 1.0 (perfect accuracy). Using criteria established by Swets (1992), the resulting value was classified as excellent ($\geq .90$), good (.80 to .89), fair (.70 to .79), or poor ($< .70$).

Q5: How does teacher acceptability of HD passages compare to acceptability of other Grade 1 reading screening measures? Means, standard deviations, and ranges were calculated for both forms of the ARP-R (HD passage survey and NWF survey). Teacher acceptability was examined by calculating a total score for each ARP-R completed. In order to calculate a total score, item responses (#1-12) were summed for both the HD passages and the DIBELS NWF surveys each teacher completed. A *t*-test was conducted to determine whether the difference

between scores was significant. A *p*-value of .05 or less was required to determine significance. Results will are also described qualitatively.

Chapter IV: Results

Preliminary Analyses

Descriptive statistics indicated total of 12 students were missing data from the spring HD passage assessment period due to attrition (moved out of the area), while two more were missing GRADE assessment data for the same reason. Due to students moving and being absent on assessment days, 18 students had missing winter DIBELS assessment data, and 19 students were missing DIBELS scores for the spring assessment; the schools the students attended supplied these data. Students missing HD passage, DIBELS, or GRADE data were excluded from the relevant correlational and cross-tabulation analyses, per conventional guidelines (Leong & Austin, 2006).

Visual inspection of histograms revealed a positive skew for scores on HD passages administered in the winter, during the retest period, and in the spring. In terms of winter DIBELS assessments, all assessments (NWF CLS, NWF WWR, and DORF WC) demonstrated a positive skew, with this being most pronounced for NWF WWR and DORF WC. By the spring, the NWF CLS distribution had normalized, although a spike in scores was evident at the positive end of the distribution. A plateau-shaped distribution was evident in the spring NWF WWR data, with most scores continuing to be concentrated between 0 and 20. Spring DORF WC data demonstrated a continued positive skew. Visual inspection of the histogram for GRADE standard scores, administered in the spring, indicated a normal distribution.

Descriptive statistics were also consulted for skewness and kurtosis values, which fell within acceptable limits (-2 to +2) for all HD passage, DIBELS, and GRADE standard scores used in analyses (Lomax, 2001). Table 1 includes additional information regarding descriptive statistics for the data collected as part of this study. It should be noted that skewness and

kurtosis statistics fell outside of the recommended range for HD passage accuracy values. However, these values were not used in the analyses conducted as part of the current study, and are included purely for qualitative information. Normal probability plots (P-P plots) were examined for univariate normality, and none of the measures (HD passage scores, DIBELS scores, or GRADE standard scores) showed any major departure from normality. Bivariate scatterplots for HD passage scores in the winter, re-test, and spring assessment periods, as well as DIBELS and GRADE standard scores were examined for bivariate normality. Scatterplots indicated linear relationships between each pair of variables and did not indicate the presence of any outliers that could influence correlation statistics.

In reviewing descriptive statistics for the HD passages and DIBELS measures, it is evident that mean words read correct per minute are similar for HD passages and DIBELS Oral Reading Fluency at the winter and spring benchmark assessments. Students tended to read slightly more words correct on the HD passages than on DIBELS passages. Moreover, students tended to make more errors when reading DIBELS passages than HD passages, as is indicated by accuracy results for these assessments. Overall, accuracy tended to be higher by 4-5% for students when they read HD passages than when reading DIBELS passages.

Analysis of Research Questions

Q1: What is the reliability (i.e., inter-rater, test-retest, and alternate form reliability) of HD passages? Inter-rater reliability was calculated for 10% of all HD passages (randomly selected at each assessment) using word-by-word agreement [agreements/(agreements + disagreements)]. Results of word-by-word agreement calculations indicated overall agreement of 98.2% across all HD passage assessments (winter, re-test, spring). At the winter assessment, agreement was 97.5%, at the re-test assessment period, agreement was 99%, and at the spring

assessment period, agreement was calculated at 98.2%. These results suggest high inter-rater reliability.

Table 2 includes results of correlations conducted to examine reliability. Results indicate strong test-retest and alternate form reliability. Specifically, Pearson Product Moment correlations conducted to examine test-retest reliability for a subset of students (96) indicated a strong relationship between the two scores ($r(94) = 0.98$, $p < 0.01$). Descriptive statistics also demonstrate similar overall scores for students from the retest group assessed in the winter, and then again four weeks later. Specifically, the 96 students selected for the test-retest analyses read an average of 52.3 words correct at the winter assessment, and then an average of 58.3 words read correct approximately four weeks later. These mean scores suggest minimal growth over the four week period between assessments, and, therefore, similar overall performance.

Pearson Product Moment correlations conducted to examine alternate form reliability also indicated the relationships between HD passage scores in the winter (HD passage 1 vs. HD passage 2: $r(232) = 0.97$, $p < 0.01$; HD passage 1 vs. HD passage 3: $r(232) = 0.97$, $p < 0.01$; HD passage 2 vs. HD passage 3: $r(232) = 0.97$, $p < 0.01$), during the retest period (HD passage 1 vs. HD passage 2: $r(94) = 0.98$, $p < 0.01$; HD passage 1 vs. HD passage 3: $r(94) = 0.98$, $p < 0.01$; HD passage 2 vs. HD passage 3: $r(94) = 0.98$, $p < 0.01$), and in the spring (HD passage 1 vs. HD passage 2: $r(220) = 0.96$, $p < 0.01$; HD passage 1 vs. HD passage 3: $r(220) = 0.96$, $p < 0.01$; HD passage 2 vs. HD passage 3: $r(220) = 0.97$, $p < 0.01$) were very strong.

Q2: What is the convergent validity (i.e., concurrent and predictive validity) of highly decodable (HD) passages for assessing reading ability in 1st grade with currently available measures of reading ability (e.g., DIBELS Next NWF, DORF, GRADE)? Strong relationships were also found between HD passage scores and currently available measures of reading ability

(see Table 3). With respect to convergent validity, Pearson Product Moment correlations suggested strong concurrent validity, as indicated by relationships between winter median HD passage scores and winter DIBELS NWF CLS ($r(214) = 0.84$, $p < 0.01$), NWF WWR ($r(214) = 0.82$, $p < 0.01$), and DORF ($r(214) = 0.96$, $p < 0.01$). In the spring, relationships between median HD passage scores and these measures were also very strong (DIBELS NWF CLS: $r(213) = 0.83$, $p < 0.01$; DIBELS NWF WWR: $r(213) = 0.81$, $p < 0.01$; DORF: $r(213) = 0.96$, $p < 0.01$), as well as correlations between spring median HD passage scores and the GRADE Comprehension Composite standard score ($r(218) = 0.84$, $p < 0.01$). Similarly, results indicated strong predictive validity when considering relationships between winter median HD passage scores and spring DIBELS scores (DIBELS NWF CLS: $r(213) = 0.82$, $p < 0.01$; DIBELS NWF WWR: $r(213) = 0.80$, $p < 0.01$; DORF: $r(213) = 0.93$, $p < 0.01$), as well as relationships between winter median HD passage scores and spring GRADE Comprehension Composite standard scores ($r(218) = 0.83$, $p < 0.01$).

Q3: How well do students' fall HD passage scores predict their winter HD passage scores, and how does inter-classroom variability affect this relationship between winter and spring scores? The unconditional model (i.e., HLM without any Level 1 or Level 2 predictors), otherwise known as a one-way ANOVA with random effects, was run to calculate interclass correlation coefficient (ICC). ICC was calculated using output for variance among students within classrooms ($\sigma^2 = 610.08$) and variance between classrooms ($\tau_{00} = 1136.40$). It was found that 39% of the variance [610.08/(610.08+1136.40) = 0.39] was due to classroom variability. With such a substantial ICC, this variability would be inappropriately ignored in OLS regression. As a result, the full, 2-level HLM was conducted, and variables at the student (winter HD median score) and classroom levels (percent of students receiving free/reduced lunch, years of teaching

experience, teacher gender, and teacher ethnicity) were added. The model is mathematically given as:

Level 1 (student level; within classroom):

$$y_{ij} = \beta_{0j} + \beta_{1j}WinterHDPScore_{ij} + r_{ij} ,$$

Level 2 (classroom level):

$$\beta_{0j} = \gamma_{00} + \gamma_{01}PERCFRLU_j + \gamma_{02}YRSTEACH_j + \gamma_{03}GENDER_j + \gamma_{04}ETHNIC_j + \upsilon_{0j} ,$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}PERCFRLU_j + \gamma_{12}YRSTEACH_j + \gamma_{13}GENDER_j + \gamma_{14}ETHNIC_j + u_{1j} ,$$

Combined, the 2-level HLM is given by:

$$y_{ij} = \gamma_{00} + \gamma_{01}PERCFRLU_j + \gamma_{02}YRSTEACH_j + \gamma_{03}GENDER_j + \gamma_{04}ETHNIC_j + \gamma_{10}W\_HD\_MED_{ij} + \gamma_{11}PERCFRLU_jW\_HD\_MED_{ij} + \gamma_{12}YRSTEACH_jW\_HD\_MED_{ij} + \gamma_{13}GENDER_jW\_HD\_MED_{ij} + \gamma_{14}ETHNICIT_jW\_HD\_MED_{ij} + u0j + u_{1j}W\_HD\_MED_{ij} + r_{ij} .$$

Fixed effects for the 2-level HLM are presented in Table 4. As can be seen, none of the classroom level variables (percent of students receiving free/reduced lunch, years of teaching experience, teacher gender, and teacher ethnicity) were significantly associated with either spring HD passage median scores or slope. Table 5 offers random effects for the 2-level HLM, which indicate a significant association between classroom membership and spring HD passage median score ($p < .05$), but no significant effect of classroom membership on slope, or growth trajectory. These significant random effects confirm the appropriateness of HLM over OLS regression, given that it is inappropriate to ignore classroom variability in this case. However, the existence of significant random effects indicates that there remains significant variation between classroom spring HD passage median scores that cannot be explained by the current model and its component predictors.

Q4: What is the diagnostic accuracy (i.e., sensitivity, specificity, negative and positive predictive power, negative and positive likelihood ratios, post-test probabilities, area under the curve) of HD passages for 1st grade students? Student outcomes on the GRADE

Comprehension Composite were converted to a binary variable (0 = no indication of reading difficulties/performance at or above the 25th/40th percentile; 1 = indication of reading difficulties/performance below the 25th/40th percentile) in order to run logistic regression using a default cut value of .5. Overall, 70 (32%) of the 220 students administered the GRADE Comprehension Composite in the spring fell below the 25th percentile, while 85 (39%) fell below the 40th percentile. Cross tabulations revealed the overall classification accuracy to be .88 when predicting both performance below the 25th percentile and below the 40th percentile, meaning that 88% of students were correctly identified as being either "at risk" or "not at risk" based on their winter HD passage scores, regardless of which cut point for risk was used.

Sensitivity and specificity of HD passages were found to be quite high. When predicting performance below the 25th percentile sensitivity was .86, meaning that 86% of students who went on to fall below the 25th percentile on the GRADE were correctly identified as "at risk" by the HD passages administered in the winter. Conversely, specificity of .89 suggests that 89% of students who performed at or above the 25th percentile were correctly identified as "not at risk." Using the 40th percentile as the cut point for determining reading difficulty, sensitivity was .89, while specificity was .87. That is, 89% of students who went on to fall below the 40th percentile on the GRADE Comprehension Composite in the spring of first grade were correctly identified by their HD passage score as being "at risk" in the winter. Conversely, the measure correctly identified 87% of those students who did not go on to have reading difficulties in the spring.

Positive and negative predictive values were also found to be quite high. Specifically, when predicting performance above or below the 25th percentile on the GRADE, 79% of the students predicted to fall below the 25th percentile were, in fact, observed to fall below this cutoff on the GRADE Comprehension Composite in the spring (true positives). Negative predictive

value indicated that 93% of students predicted to perform at or above the 25th percentile on the GRADE were observed to do so at that time. When predicting performance above or below the 40th percentile, positive predictive value was 82%. That is, 82% of those students predicted to fall below the 40th percentile were observed to fall below this cutoff on the GRADE Comprehension Composite (true positives). In terms of negative predictive value, 93% of students predicted to perform at or above the 40th percentile based on winter HD passage scores were observed to perform at this level in the spring (true negatives).

Results of logistic regression indicated that the model including winter HD passage scores made a significant improvement to model fit over the empty model with the intercept only at step 0 when predicting to either the 25th percentile ($\chi^2_{(1)} = 163.51$, $p < .001$) or 40th percentile ($\chi^2_{(1)} = 179.02$, $p < .001$) on the spring GRADE Comprehension Composite.

Following steps outlined by VanDerHeyden (2010), positive and negative likelihood ratios were identified. Positive likelihood ratios of 8.01 and 7.10 were identified when predicting to performance above or below the 25th and 40th percentiles respectively. Negative likelihood ratios of 0.16 and 0.12 were identified when predicting to the 25th and 40th percentiles respectively. These ratios were used to compute post-test probabilities. In terms of predicting performance above or below the 25th percentile, a positive post-test probability of 79% was identified, while a negative post-test probability of 7% was found. That is, the probability was 79% that a student identified as "at risk" by the HD passage screening measure in the winter would be identified as a struggling reader in the spring (i.e., obtaining a score below the 25th percentile on the GRADE Comprehension Composite). Conversely, there was a 7% chance that a student identified as "not at risk" by the HD passage winter screening would go on to have reading difficulties.

In terms of predicting performance above or below the 40th percentile, a positive post-test probability of 82% was identified, while a negative post-test probability of 7% was found. That is, the probability was 82% that a student identified as "at risk" by the HD passage screening measure in the winter would be identified as a struggling reader in the spring (i.e., obtaining a score below the 40th percentile on the GRADE Comprehension Composite). Conversely, there was a 7% chance that a student identified as "not at risk" by the HD passage winter screening would go on to have reading difficulties.

Finally, results of ROC curves analyses indicated excellent classification accuracy with a high AUC estimate of .95, $p < .001$, 95% CI = .93–.98 (using performance below the 25th percentile on the GRADE to define reading difficulties) 96, $p < .001$, 95% CI = .93–.98 (using performance below the 40th percentile on the GRADE to define reading difficulties). These results can be interpreted as indicating a 95% likelihood that students earning a score at or above the 25th percentile on the GRADE Comprehension Composite will have a higher winter HD passage score than those students who earned scores below the 25th percentile. Likewise, there is a 96% likelihood that students earning a score at or above the 40th percentile on the GRADE Comprehension Composite will have a higher winter HD passage score than those students who earned scores below the 40th percentile. Further evidence of these results is offered through an examination of plots of the ROC curves above the reference line, which indicates chance, and represents an AUC of 0.5 (see Figures 1 and 2).

Q5: How does teacher acceptability of HD passages compare to acceptability of other Grade 1 reading screening measures? Means, standard deviation, and ranges for the ARP-R were analyzed first. Results indicated nearly identical mean scores for the HD passage survey ($M = 46.1$, $SD = 16.0$) and NWF survey ($M = 46.2$, $SD = 15.0$), while ranges were also very

similar (HD passage survey score range: 12–68; NWF survey score range: 12–62). Teacher acceptability total scores were derived by summing the linear combination of responses to items 1 through 12 of the ARP-R. Results of a dependent/paired samples *t*-test indicated a nonsignificant effect of measure type ($t(20)$ -0.01, $p = .99$). That is to say, teachers rated the HD passage measure and NWF measure as similarly acceptable, overall.

Chapter V: Discussion

To psychometrically evaluate a new early reading screening measure (Highly Decodable Passages), 234 first grade students in 15 classrooms were administered the HD passages in the winter, and again in the spring. A randomly selected group of 96 students were also assessed during a re-test period approximately 4 weeks after the winter assessment. As an external criterion outcome measure, all participating students were also administered the GRADE Comprehension Composite in the spring. Schools supplied DIBELS screening data for the winter and spring benchmark periods, and a group of 20 teachers completed acceptability surveys to compare teacher views on a nonsense word fluency measure versus the HD passages. Results of reliability, validity, and predictive utility analyses offer very promising results for the newly developed HD passages. Teacher acceptability data makes it unclear whether HD passages might serve as a more acceptable screening measure to teachers than some other currently available measures, such as NWF.

In terms of descriptive statistics, HD passages reflected growth over time, with students reading fewer words in the winter (M = 47.6 WRC) than in the spring (M = 66.6 WRC). Furthermore, accuracy improved from an average of 83% in the winter, to an average of 90.9% in the spring. Compared to DIBELS ORF passages (winter M = 78.9%; spring M = 86.6%), students demonstrated between 4% and 5% greater accuracy on HD passages. It is important to note this difference because of the potential impact the experience of reading these passages may have on students. As early readers, students require motivation to continue learning to read. If we consider that school should function as an environment that encourages learning, our assessments should also foster motivation in our young readers. The experience of successfully reading assessment passages is surely to motivate students more than the experience of reading

difficult assessment passages that include a larger number of difficult words that a student is unable to read correctly.

Reliability Results

Reliability analyses indicate the HD passages provide a highly reliable approach to measurement. When considering inter-rater reliability, agreement between different raters was extremely similar, with average agreement ranging from 97.2% at the winter assessment to 99% at the spring assessment. These results are consistent with those for other curriculum-based measures of early literacy and reading ability (e.g., Clemens, Shapiro, & Thoemmes, 2011). Similarly, results of correlational analyses indicate strong test-retest and alternate form reliability. Test-retest reliability for a subset of 96 students tested approximately 4 weeks after the winter assessment revealed a strong relationship between scores ($r(94) = 0.98$, $p < 0.01$). This suggests that HD passages can be administered with confidence that a student's score will be consistent within a short period of time.

It is important to note that the test-retest interval in this study was longer than typically adopted. Indeed, the interval for the current study was originally planned to be only 2 weeks. However, because of winter break in late December, it was not possible to test students as quickly as intended. It is interesting to note that despite a 4-week interval between winter assessments and retesting for the 96 students in the test-retest sample, scores remained relatively stable, resulting in a high test-retest correlation. This is likely a result of the fact that students did not receive instruction for approximately 2 weeks while out of school on vacation, which may have improved test-retest results for the current study. This is because students were not experiencing a great deal of growth during the test-retest interval, therefore improving the stability of scores across time.

Alternate form reliability suggests that different HD passages will yield consistent results for students regardless of which passage is administered, as indicated by strong correlation coefficients between each of the three passages administered at each of the three assessment points (range = .96-.98). This represents a similarly high level of reliability compared to alternate form reliability reported for DIBELS NWF CLS (r = .85) and NWF WWR (r = .90; Good et al., 2013). Single form alternate reliability is not reported in the DIBELS Next Technical Manual (Good et al., 2013). Alternate form reliability using the median of 3 passages is reported to be between .97 and .98. Alternate form reliability of AIMSweb R-CBM passages for first grade students has been demonstrated to range from .94 to .95 at each benchmark assessment (Pearson, 2012a), which is comparable to that of HD passages in the current study. In general, these results are very promising and indicate that HD passages are an extremely reliable measure of reading ability.

Convergent Validity Results

Similarly, HD passages demonstrated extremely high levels of convergent validity, as indicated by correlations with other currently available measures of early reading ability. In terms of concurrent validity (for assessments administered at the same point in time), correlation coefficients ranged from .81 (for spring HD passages and DIBELS NWF WWR) to .96 (for winter and spring HD passages and DORF Words Correct). With respect to predictive validity, relationships were also strong for all measures, with correlations ranging from .80 (for winter HD passages and spring DIBELS NWF WWR) to .93 (for winter HD passages and spring DORF Words Correct). It is evident from these values that HD passage scores are strongly related to other early reading screening measures, regardless of whether they are administered concurrently, or at different points in time. In particular, HD passages are strongly related to other measure of

oral reading (i.e., DORF). This may offer additional support for the validity of HD passages, as curriculum-based measures of oral reading demonstrate the most extensive evidence as a reading screening measure (e.g., Ball & Christ, 2012).

In addition, using the GRADE Comprehension Composite (administered in the spring), concurrent and predictive validity was most impressive for HD passages. Concurrent validity was .84 for HD passages in the spring, which was matched only by the spring administration of DIBELS ORF. Correlations with other spring DIBELS measures were lower, including DIBELS NWF CLS (.71) and DIBELS NWF WWR (.73). Predictive validity was somewhat weaker for these measures, falling in the moderate range for DIBELS NWF CLS (.65) and DIBELS NWF WWR (.67). Predictive validity was higher for DIBELS ORF when using the GRADE Comprehension Composite as a spring outcome measure (.78). Winter HD passage scores demonstrated the strongest correlation with the spring GRADE administration (.83). These results suggest HD passages offer a valid indication of a student's reading abilities, as measured by several tests of reading proficiency, including an assessment of reading comprehension. Strong relationships between HD passages and other screening measures, as well as standardized criterion outcome measures are evident.

HLM Results

Results of hierarchical linear modeling indicated a relationship between winter and spring HD passage scores, with substantial classroom variability (39% variance attributable to classroom variability). Although there was clear variability in the spring HD passage score based on classroom membership, the classroom level variables included in the 2-level HLM could not explain this variability. Classroom level variables included percent of students receiving free/reduced lunch, years of teaching experience, teacher gender, and teacher ethnicity.

Given the preponderance of evidence suggesting the impact of socioeconomic status on achievement (e.g., Sirin, 2005), it is surprising that the variable for this (percent of students receiving free/reduced lunch) was not significantly associated with either HD passage median scores or slope. Similarly, it is surprising that characteristics of the teacher, such as years of experience, gender, and ethnicity had no significant impact on the intercept or slope. In particular, years of teaching experience is one teacher characteristic that has been demonstrated to have an effect on student achievement (e.g., Rice, 2011). However, this variable did not appear to be significantly associated with either spring HD passage scores or rate of growth from winter to spring.

There are several reasons why none of the classroom variables were significantly associated with either intercept or slope. First of all, it may be that the sample size was not large enough to detect significant effects. Unfortunately, the sample did not reach the recommended minimum outlined by de Leeuw and Kreft (1995). Rather than the recommended minimum of 20 groups (classrooms), there were only 15 in the current study. Another possible explanation for the absence of significant results may be that many classrooms shared a common school environment, which may have reduced the variability between the classroom groups and therefore made significant results difficult to detect. In this case, there were four schools in the study, where between 3 and 5 participating first grade classrooms were located. Classrooms within schools shared a common curriculum, approach to providing academic support, and school culture, which is likely to have reduced variability between classrooms in those schools. Unfortunately, there were not enough schools in the present study to conduct HLM using schools as a grouping variable. Finally, two of the variables, teacher gender, and teacher race, included very limited variability, with 14 of the 15 teachers being female, and 14 of the 15 teachers being

White. With such restricted variability in these classroom-level variables, significant outcomes will be difficult, if not impossible to detect.

It seems most likely that there are aspects of the classroom, such as socioeconomic make-up of the students and instructional characteristics common to classrooms within a school that would influence the spring HD passage scores as well as the growth from winter to spring that students demonstrate. Unfortunately, the small sample size and limited classroom variability negatively impacted our ability to effectively investigate these variables. Given research on effective instruction, it is plausible that classrooms and schools where teachers have received training and support in using research-based approaches to reading instruction will see first grade students with greater performance on HD passage assessments (e.g., Foorman & Moats, 2004; Podhajski et al., 2009; Spear-Swerling, 2009).

Diagnostic Accuracy Results

Results of the current study suggest that HD passages offer a promising new approach to screening, one that demonstrates high sensitivity and specificity. In particular, not only was the winter HD passage assessment able to correctly predict 86% and 89% of those students who went on to have reading difficulties (performed below the 25th and 40th percentile on the GRADE Comprehension Composite respectively), but it was also able to correctly predict 89% and 87% of those first graders who went on to have *no* reading difficulties (performed at or above the 35th and 40th percentiles respectively). This high level of *both* sensitivity and specificity is uncommon, and suggests that HD passages are capable of effectively discriminating between readers who are and are not "at risk" without over identifying students as "at risk" or missing a great deal of those who are.

According to Jenkins, Hudson, and Johnson (2007), an effective screening measure should demonstrate sensitivity of at least .90, with the highest specificity possible. In the current study, the default cut score of 0.5 was used for logistic regression, which yielded sensitivity of .86 and .89 and specificity of .89 and .87. It is possible to set sensitivity to .90 or higher, which would then reduce specificity somewhat. This is a trade-off that must be weighed by researchers and educators in setting the cut point for application in schools. Considering that the sensitivity demonstrated by HD passages when predicting performance below the 25th or 40th percentile in the current study is still quite close to the recommended minimum (.90) while maintaining a very high degree of specificity, one may not want to change the cut point and alter these values, as the efficiency for schools in avoiding false positives offers a resource-saving advantage. Using the default cut point of 0.5, overall classification accuracy was also very high, at .88 when using either the 25th or 40th percentile to define the cut point for reading difficulty. That is, 88% of students were correctly classified as either "at risk," or "not at risk" in the current study. As a comparison, in a study by Johnson, Jenkins, Petscher, and Catts (2009), the authors found that overall classification accuracy was only 41% for PSF and 57% for NWF in first grade when sensitivity was set to .90. These findings, again, offer an impressive picture of the potential of HD passages as an early reading screening measure.

Furthermore, these findings are quite impressive when one considers the base rate of students with reading difficulty in the sample. Specifically, 32% of students in the sample performed below the 25th percentile on the GRADE, while 39% performed below the 40th percentile. Given these statistics, if one were to designate all students as "not at risk," we would be left with an error rate of 32% and 39%. Following reasoning outlined by VanDerHeyden and Witt (2005) in their study of the accuracy of a problem-solving model for identification, the goal

for HD passages would be to surpass an accuracy rate of 68% and 61% when predicting performance below the 25th and 40th percentiles on the GRADE respectively. As can be seen, using winter HD passage scores to predict reading difficulty is clearly superior than assuming no students will go on to have reading difficulties. Specifically, with an accuracy rate of 88% and an error rate of just 22%, HD passages clearly offer a distinct advantage over chance prediction.

Effective screening measures should also demonstrate acceptable likelihood ratios and post-test probabilities as further evidence of their ability to correctly identify students beyond chance prediction when considering base rates (VanDerHeyden, 2010). Specifically, positive likelihood values are indicative of an effective screening measure when they are higher than 1.0, which would indicate equivalent pre-test and post-test probabilities. In this case, we see that HD passages demonstrate positive likelihood ratios of 8.01 and 7.10 (when predicting performance below the 25th and 40th percentiles respectively), which are much greater than 1.0, and quite large for a sample with the prevalence of reading difficulty (performance below the 25th/40th percentile on the GRADE) at 31.8% and 38.6%. Furthermore, the negative likelihood ratios are quite low (.16 and .12), suggesting a strong likelihood that a student predicted to be "not at risk" by winter HD passage performance will go on to demonstrate success on the GRADE (i.e., no reading difficulties will be evident). According to VanDerHeyden (2010), "higher [positive] likelihood ratios (e.g., on the order of 10 or better) are required for the test to improve over chance prediction where prevalence is very high (e.g., 70%)." In this case, positive likelihood ratios approached a value of 10 despite a much lower prevalence rate (32% and 38% depending on the cut point for determining reading difficulty).

These findings are further supported by post-test probabilities. In terms of post-test probabilities for falling below the 25th percentile on the GRADE Comprehension Composite (i.e.,

demonstrating reading difficulties), there is an 79% probability that a student who was identified as "at risk" according to the winter HD passage score would go on to demonstrate reading difficulties. Conversely, there was only a 7% chance that a student predicted to demonstrate reading success on the GRADE would actually demonstrate reading difficulty. When predicting performance below the 40th percentile on the GRADE Comprehension Composite, there is an 82% probability that a student who was identified as "at risk" would go on to demonstrate reading difficulties, and a 7% chance that a student predicted to demonstrate reading success on the GRADE would actually demonstrate reading difficulty. These statistics clearly demonstrate that the HD passages offer substantial utility beyond chance prediction, and can assist educators in discriminating between students who are in need of intervention and those who are not.

Additional evidence of diagnostic accuracy is offered from the ROC curves analysis. AUC estimates of .95 and .96 fell well within the excellent range of classification accuracy (Swets, 1992). This is quite impressive, indeed, considering findings from other studies investigating AUC for various screening measures. For instance, in a study of third grade students DORF AUC was found to be .82 when predicting end of year state test performance from December of third grade (Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008), a time when R-CBM is considered a highly effective screening measure (Ball & Christ, 2012). In the earlier grades, when predicting DORF performance at the end of first grade, DIBELS NWF demonstrated the highest AUC of all kindergarten and first grade screening measures, at .85, while LNF, ISF, and PSF values were .82, .70, and .65 respectively. Taken together, these results demonstrate that the diagnostic accuracy of HD passages is extremely high. Furthermore, HD passages may actually be superior to many existing early screening measures in their ability to identify those students in need of reading interventions, while helping schools to conserve

valuable resources by avoiding the unnecessary identification of students who do not require intervention.

When comparing the diagnostic accuracy results for HD passages when using performance below the 25th percentile versus performance below the 40th percentile on the GRADE Comprehension Composite to define reading difficulty it is difficult to say which approach works better for screening. It is clear that HD passages retain nearly identical psychometric strengths when predicting to either outcome, offering evidence of its robustness to changes in base rate of reading failure and definition of reading failure. The fact that HD passages perform essentially equally when predicting to either outcome suggests that schools will have the luxury of selecting which approach works best for them, or using both. However, given that screening measures are intended to identify students in need of supports and to provide these to students before students fall too far behind, schools may to do well to adopt a more conservative approach and identify more children in need of academic support by selecting performance below the 40th percentile as the cut point for determining "at risk" status. At the same time, some schools may not have the resources to devote to more students than absolutely necessary, or may have other reasons for selecting prediction to the 25th percentile over the 40th. Either way, these results suggest that HD passages offer a viable approach to screening that is robust to changes in the definition of risk status.

Teacher Acceptability Results

Results of the acceptability survey failed to indicate whether teachers may prefer HD passages to existing assessments of decoding that rely on nonsense words. The sample size for this analysis fell short of the required minimum number of teachers. However, means for surveys completed on each measure (HD passage = 46.1, NWF = 46.2) were nearly identical,

suggesting that the inability of the *t*-test to find significant results was not solely an issue of sample size. It is possible, however, that with a larger sample of teachers, differences in survey results would have become apparent. Looking qualitatively at the results, it was clear that teachers often had strong opinions about each approach to assessment, either feeling that nonsense words were insufficient for assessing first grader readers *or* feeling decodable text passages were insufficient. Overall, however, these polarized opinions balanced one another out in the end, with mean scores being nearly identical. The landscape of responses does suggest that educators have not universally accepted existing screening protocols, and that opportunity exists to introduce more effective, socially valid approaches to measurement.

Limitations

It is important to consider the limitations of this study when interpreting the results. First, one must consider that current study's focus is on students in the middle and end of first grade. Unfortunately, due to time constraints and other logistical issues, it was not possible to assess students in the beginning of first grade. This leaves the question of whether HD passages may offer a viable alternative to existing measures at a point in time when the most serious psychometric and practical limitations are evident. It is likely that this study has identified a middle or end point of a window when HD passages are effective as a screening measure. As research suggests, existing R-CBM passages offer psychometrically and practically efficient screening tools as students enter the middle and upper grades (e.g., Ball & Christ, 2012). Furthermore, the current study suggests that HD passages *may* offer only slight improvement over grade 1 DIBELS passages, as indicated by strong correlations and similar mean words read correct. Unfortunately, the current study does not answer the question of how early HD passages are effective with students. It may be that HD passages are effective at detecting developing

reading skills earlier than DORF passages and other available screening measures, although this particular question was not able to be addressed in the current study. This is a limitation of the current study, and one that should be addressed through future research.

An additional consideration that must be made is that participating students came from four schools in Eastern Pennsylvania, and represent a restricted geographic region. Additionally, three of the schools came from the same school district, further restricting the sample of students and teachers. As a result, caution must be taken when generalizing results to other students in other regions, and additional research will be needed to replicate results. With a restricted sample size in terms of number and diversity, especially geographic, the results of this study must be interpreted with caution. One cannot be confident that results would be replicated across different geographic regions, in school districts with different racial/ethnic and socio-economic make-ups, or in schools where the instruction varies from those in the current study. Future research must aim to replicate findings of reliability, convergent validity, and diagnostic accuracy in order to confirm the psychometric strengths of the HD passages.

Furthermore, results of HLM and the acceptability study should be considered with caution as a result of the small sample sizes for these analyses. Specifically, while a minimum number of 20 groups (classrooms) are recommended (de Leew & Kreft, 1995), only 15 classrooms were included, and these classrooms were nested within four schools. Classrooms in the same school shared a curriculum, as well as plans for providing students with intervention who required additional support. For instance, one school grouped students across the grade by level of need, and assigned each group to a specific teacher or reading specialist for small group support at various times during the grade's ELA block. This greatly reduced the variability in instruction and interaction between classrooms, as students from different classrooms were

combined into small groups for reading instruction. A larger sample size with enough schools to analyze variables at the school level would be more effective for examining aspects of the environment that might influence HD passage intercept and slope.

Similarly, a sample size of 20 teachers fell below the required 34 to conduct dependent-means *t*-test on acceptability survey results. This small sample of surveys yielded similar mean scores for both the NWF survey and the HD passages survey. With such a small group of teachers, it is difficult to determine if these responses offer a representative reflection of educators' views of these measures. Additionally, when looking closely at the data, it is evident there is no consistent pattern to responses. While some teachers view both measures equally acceptable, others strongly prefer one measure over the other. Additional responses are required to more fully understand this question.

Implications for Practice

With consideration for these limitations, this psychometric study of the newly developed HD passages offers promising implications for research. In terms of its scientific properties, the HD passages appear to be a highly reliable and valid measure for screening early readers at the middle and end of first grade. Although existing measures intended for kindergarteners and first graders present with a host of problems, including low correlations with future reading measures, and poor diagnostic accuracy (e.g., Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009; Johnson, Jenkins, Petscher, & Catts, 2009), HD passages demonstrate high levels of concurrent and predictive validity, as well as a high AUC value, and impressive sensitivity, specificity, and post-tests probabilities. This suggests that HD passages may offer distinct advantages over existing screening measures in terms of its scientific properties.

Should future studies confirm these preliminary findings, this also suggests that practical advantages will be provided to schools. For instance, with a measure that demonstrates such strong scientific properties, it is unlikely that it will be necessary to administer additional measures as part of the screening process at earlier grades. This is extremely important to schools, where instructional time is at a premium, and even a few more minutes of assessment time adds up quickly when it must be devoted to hundreds of students. Additionally, the strong psychometric properties of the HD passages may allow schools to more precisely identify students who are truly in need of academic intervention and supports. In particular, specificity, or the ability to correctly identify students who *will not* go on to experience reading difficulty, has often been sacrificed for higher levels of sensitivity, or the ability to correctly identify students who *will* go on to experience reading difficulty. In the case of HD passages, it appears that educators will be able to place greater confidence in the fact that students who are identified as needing intervention truly need it, rather than providing intervention to many students who do not truly need it as a result of screening measures that yield a high number of false positives.

Furthermore, HD passages offer an advantage over DIBELS ORF passages in that they appear less difficult for students to read, as indicated by a higher degree of accuracy (between 4% and 5% higher for HD passages). When considering that students in an educational context, and especially an assessment context, should be motivated to do their best work, this improvement should not be overlooked. The benefits of using passages that are less difficult for students, especially in a testing situation, may include the fact that students will be more motivated to perform if they are experiencing fewer difficult words. A match between ability and assessment material is an important consideration in the development of new screening tools. The results of the current study suggest that HD passages may offer a better fit than existing

measures, which may lead to a more accurate picture of student skills as a result of their increased motivation and effort with assessment material.

Future Research Directions

Results of the current study offer valuable directions for future research. An essential question that was not addressed in the current study is how early HD passages can be effectively used as a screening measure, and whether these passages are able to identify reading abilities earlier than existing measures. As discussed previously, the current study clearly suggests that HD passages are effective in identifying students in the middle of first grade who will go on to demonstrate reading difficulties at the end of the year. However, considering limitations of earlier measures for kindergarten and first grade students, this does not fully answer the question of when HD passages can begin to be used and whether they can replace existing early reading screeners. Future research should aim to address these questions by evaluating HD passages at progressively earlier points in reading development. For instance, do HD passages demonstrate similar levels of reliability, validity, and predictive accuracy when used at the beginning of the first grade year or at the end of kindergarten? By expanding research to include a wider span of time researchers can begin to identify the window within which HD passages are most effective, and whether or not these passages can be used to replace existing early reading screening measures.

Moreover, in addition to establishing the earliest point at which HD passages can be effectively used as a screening measure, future research should investigate whether HD passages can be utilized as a progress monitoring tool. Does regularly collected data using HD passages offer valid and reliable information for making informed instructional decisions? How frequently must educators assess students using HD passages when monitoring progress? How

old should students be for HD passages to act as an effective progress monitoring tool? Related to investigating HD passages with younger students and expanding research to evaluate their utility as a progress monitoring tool, future research should also aim to compare practical and psychometric properties of HD passages directly to those of existing early literacy CBMs. For instance, researches should directly compare the sensitivity, specificity, and AUC values of HD passages versus measures such as DORF and DIBELS Next NWF.

Furthermore, it is important that results be replicated with larger samples of students in a wider geographic area, and with greater diversity. In particular, HLM should be repeated with a much larger sample that includes a large number of schools from different districts in order to investigate variables at this level that may be significantly associated with intercept and slope of the HD passages. Additionally, with a larger number of schools and districts, researchers may be able to identify specific aspects of the teaching environment that may contribute to HD passage outcomes and growth over time. Potential variables for future study might include instructional practices (e.g., use of direct instruction), behavioral management (e.g., school-wide positive behavior support), or specific curricula or intervention protocols used.

It is doubtful that simply identifying a reading curriculum or behavioral management approach (e.g., existence of school-wide positive behavior support) will be specific enough to explain variance in either intercept or slope of HD passages. For example, although several schools in the current study reported using the same curriculum, each of those schools was implementing an RTI model in a slightly different way, and teachers and reading specialists were adopting components of the curriculum and other supplemental programs that were different from each other. Therefore, it might be interesting to investigate such variables by using more involved observational systems that allow researchers to code specific teacher behaviors. This

may more effectively produce variability between classrooms and schools, while also pinpointing specific practices unique to each classroom that allow students to demonstrate different levels of growth.

Future studies might also investigate whether differences exist in psychometric properties of HD passages based on student characteristics, such as ethnicity. This phenomenon, referred to as predictive bias, is one aspect of measurement that is beginning to be investigated with R-CBM. For instance, Roehrig, Petscher, Nettles, Hudson, and Torgesen (2008) conducted logistic regression to determine if DORF cut scores predicted reading difficulties equally for various groups, such as students receiving free/reduced lunch, English language learners, African American students, and Latino students. Results indicated that DORF scores predicted performance equally well for all groups, indicating an absence of predictive bias. However, another group of researchers (Pearce & Gayle, 2009) found that DORF scores accounted for different amounts of variance based on a student's ethnic group. For instance, DORF accounted for 41% of the variance in the outcome measure for American Indian students, but 37% for White students as a result of a y-intercept bias that became apparent when the authors compared the two groups' regression models. Similar investigations with HD passages would offer valuable information for determining the degree to which cut scores on the HD passages are appropriate for various groups of students.

Another vital future direction to consider is acceptability. The current study did not determine whether HD passages would offer a more acceptable alternative to currently available reading screening measures. Although the use of connected text may offer a more authentic assessment of reading ability, it is unclear whether this is enough to make this measure acceptable to teachers. Because it is so important that teachers use data from measures like HD

passages to make instructional decisions, it is vital that future research investigate whether HD passages are viewed as informative and useful to the teachers we hope will use them. Another consideration is that the acceptability in this study used case examples to present data from a hypothetical student who has been assessed using an NWF measure and HD passages. Future research might address the issue of acceptability by asking teachers about their views of these measures after each teacher has personally administered and interpreted the measure. While examples of resulting data offers a realistic idea of what a measure offers, it is not exactly the same as administering and interpreting the data with one's own students. With greater familiarity, teachers may provide acceptability data that allows researchers to better discriminate between teachers' views of various measurement approaches.

Finally, it would be interesting to take the investigation of HD passages one step further by including qualitative information about the features of a child's reading (e.g., prosody, self-corrections), or asking comprehension questions following each passage. Such additions to HD passage assessments may address concerns individuals have had about existing screening measures that these tests address only isolated skills out of context (Goodman, 2006), and that they do not address comprehension (Hoffman, Jenkins, & Dunlap, 2009). By incorporating such an addition to HD passage assessments, researchers may both increase the psychometric properties of the measure, while also increasing it's acceptability to teachers. For example, researchers might simply provide an estimate of the percent of time a student was observed engaging in various behaviors during reading, such as reading with prosody, self-correcting errors, etc. These are questions that should be considered for future research.

Conclusions

The purpose of the current study was to investigate the psychometric properties and acceptability of a newly developed screening measure for early readers, Highly Decodable Passages (HD passages). Research questions addressed issues of reliability, convergent validity, predictive utility and classroom variability, predictive accuracy, and acceptability. Results suggest that HD passages offer a promising alternative to currently available early screening measures, one that is highly reliable, is highly correlated with existing measures of reading ability, and a tool that is capable of accurately discriminating between students who will and will not go on to have reading difficulties. Small sample sizes offered limited information about how classroom variability may affect HD passage intercept and slope, and the acceptability of HD passages measure compared to existing nonsense word fluency measures.

Additional investigations with larger, more geographically diverse student and teacher samples are required, as well as replication with younger students. However, results of this study offer an important step forward in the area of early reading screening. In particular, HD passages may offer the opportunity to reduce the need to administer several measures to just one, and to utilize a measure that reflects a more authentic approach to reading assessment by incorporating connected text. In addition, this measure's success in this study demonstrates the importance of considering the complex development of early reading skills in creating universal screening measures. As Fuchs (2004) explains, an effective CBM should be psychometrically sound as a result of the fact that it includes an indicator of proficiency that requires mastery of multiple skills. HD passages appear to do just that, by employing connected text that is developmentally appropriate for these young readers. It appears that the results of this study are just the beginning of a change in the approach to screening beginning readers.

References

Alonzo, J., Tindal, G., Ulmer, K., & Glasgow, A. (2006). *EasyCBM online progress monitoring assessment system.* Eugene, OR: Center for Educational Assessment Accountability.

American Institutes for Research. (2007). *Reading First state APR data.* Washington, DC: Author.

Ardoin, S.P., Suldo, S.M., Witt, J., Aldrich, S., & McDonald, E. (2005). Accuracy of readability estimates' predictions of CBM performance. *School Psychology Quarterly, 20*(1), 1-22.

Bagnato, S.J. (2007). Authentic assessment for early childhood intervention: Best practices. New York: Guilford.

Ball, C.R. & Christ, T.J. (2012). Supporting valid decision-making: Uses and misuses of assessment data within the context of RTI. *Psychology in the Schools, 49*(3), 231-244

Catts, H.W., Petscher,Y., Schatschneider, C., Bridges, M.S., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities, 42*(2), 163-176. doi: 10.1177/0022219408326219

Children's Educational Services. (1987). *Test of Reading Fluency (TORF).* Eden Prairie, MN: Author.

Clemens, N.H., Hilt-Panahon, A., Shapiro, E.S., and Yoon, M. (2012). Tracing student responsiveness to intervention with early literacy indicators: Do they reflect growth toward reading outcomes? *Reading Psychology, 33,* 47-77. doi: 10.1080/02702711.2011.630608

Clemens, H.H., Shapiro, E.S., & Thoemmes, F. (2011). Improving the efficacy of first grade

reading screening: An investigation of word identification fluency with other early literacy indicators. *School Psychology Quarterly, 26*(3), 231-244.

Compton, D.L., Appleton, A.C., & Hosp, M.K. (2004). Exploring the relationship between text leveling systems and reading accuracy and fluency in second-grade students who are average and poor decoders. *Learning Disabilities Research & Practice, 19*(3), 176-184.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological measurement, 20*, 37-46. doi: 10.1177/001316446002000104

de Leeuw, J., & Kreft, I.G.G. (1995). Questioning multilevel models. *Journal of Educational and Behavioral Statistics, 20*(2), 171-189.

Deno, S.L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219-232.

Deno, S.L. & Marston, D. (2006). Curriculum-based measurement of oral reading: An indicator of growth in fluency. In S.J. Samuels & A.E. Farstrup (Eds.), *What research has to say about fluency instruction* (pp. 179-203). Newark, DE: International Reading Association.

Deno, S.L., Mirkin, P.K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*(1), 36-45.

Eckert, T.L. & Shapiro, E.S. (1999). Methodological issues in analog acceptability research: Are teachers' acceptability ratings of assessment methods influenced by experimental design? *School Psychology Review, 28*(1), 5-16.

Eckert, T.L., Hintze, J.M., & Shapiro, E.S. (1999). Development and refinement of a measure for assessing the acceptability of assessment methods: The Assessment Rating Profile-Revised. *Canadian Journal of School Psychology, 15*(1), 21-42. 10.1177/082957359901500103.

Edformation. (2005). *AimsWeb.* Eden Prairie, MN: Author.

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

Field, A. (2012). *Discovering statistics using IBM SPSS Statistics* (4th ed.). London, England: Sage.

Foorman, B.R. & Moats, L.C. (2004). Conditions for sustaining research-based practices in early reading instruction. *Remedial and special education, 25*(1), 51-60. doi: 10.1177/0741932 5040250010601

Fuchs, L.S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities Research and Practice, 18*(3), 172-186.

Fuchs, L.S., & Deno, S.L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57*, 488-500.

Fuchs, L.S. & Deno, S.L. (1994). Must instructionally useful performance assessment be based in the curriculum? *Exceptional Children, 61*(1), 15-24.

Fuchs, L.S., Fuchs, D., & Compton, D.L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children, 71*(1), 7-21.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989). Monitoring reading growth using student recalls: Effects of two teacher feedback systems. *Journal of Educational Research, 83,* 103-111.

Furr, M. R. & Bacharach, V. R. (2008). *Psychometrics: An Introduction.* Thousand Oaks, CA: Sage Publications.

Goffreda, C.T. & DiPerna, J.C. (2010). An empirical review of psychometric evidence for the

Dynamic Indicators of Basic Early Literacy Skills. *School Psychology Review, 39*(3), 463-483.

Goffreda, C.T., DiPerna, J.C., & Pedersen, J.A. (2009). Preventative screening for early readers: Predictive validity of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). *Psychology in the Schools, 46*(6), 539-552. doi: 10.1002/pits.20396

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

Good, R.H. & Kaminski, R.A. (2011). *DIBELS Next Assessment Manual*. Dynamic Measurement Group. Available: http://www.dibels.org/

Good, R.A., Kaminski, R.A., Dewey, E.N., Wallin, J., Powell-Smith, K.A., & Latimer, R.J. (2013). *DIBELS Next Technical Manual.* Eugene, OR: Dynamic Measurement Group.

Goodman, K.S. (2006). A critical review of DIBELS. In K.S. Goodman (Ed.), *The truth about DIBELS: What it is, What it does.* (pp. 1-39). Portsmouth, NH: Henemann.

Gough, P.B. & Tunmer, W.E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*(1), 6-10. doi: 10.1177/074193258600700104

Harcourt Assessment, Inc. (2003). *Stanford Achievement Test, 10th Edition.* San Antonio, TX: Author.

Harcourt Brace Educational Measurement. (1996). *Stanford Achievement Test* (9th ed.). San Antonio, TX: Harcourt Assessment.

Harcourt Brace Educational Measurement. (2000). *Metropolitan Achievement Test* (8th ed.). San Antonio, TX: Harcourt Assessment.

Hiebert, E.H., & Fisher, C.W. (2007). Critical word factor in texts for beginning readers. *Journal of*

Educational Research, 101(1), 3-11.

Hoover, W.A. & Gough, P.B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 2*, 127-160.

Hox, J. (2010). *Multilevel analysis techniques and applications* (2nd ed.). New York, NY: Routledge.

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), 582-600.

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.

Joshi, R.M. & Aaron, P.G. (2000). The component model of reading: Simple view of reading made a little more complex. *Reading Psychology, 21*, 85-97.

Just, M.A., & Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329-355.

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), 437-447.

Kaminski, R.A. (1992). *Assessment for the primary prevention of early academic problems: Utility of curriculum-based measurement prereading tasks.* (Doctoral dissertation). Retrieved from ProQuest. (Order Number 9238932)

Kaminski, R.A. & Good, R.H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 25,* 215-227.

Karlsen, B., Madden, R., & Gardner, E. F. (1975). *Stanford Diagnostic Reading Test*. New York: Harcourt, Brace, Jovanovich.

Karlsen, B. K., & Gardner, E. (1985). *Stanford Diagnostic Reading Test* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Karlsen, B., & Gardner, E. F. (1995). *Stanford Diagnostic Reading Test* (4th ed.). San Antonio, TX: Harcourt Assessment.

Keller-Margulis, M.A., Shapiro, E.S., & Hintze, J.M. (2008). Long-term diagnostic accuracy of curriculum-based measures of reading and mathematics. *School Psychology Review, 37*(3), 374-390.

Kendeou, P., Savage, R., & van den Broek, P. (2009). Revisiting the simple view of reading. *British Journal of Educational Psychology, 79,* 353-370. doi: 10.1348/978185408X369020

Klauda, S.L.& Guthrie, J.T. (2008). Relationships of three components of reading fluency to reading comprehension. *Journal of Educational Psychology, 100*(2), 310-321. doi: 10.1037/0022-0663.100.2.310

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

Leong, F. T L., & Austin, J. T. (2006). *The psychology research handbook: A guide for graduate students and research assistants.* Thousand Oaks: Sage Publications.

Lomax, R. G. (2001). *Statistical concepts: A second course for education and the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Lopez, F.A., Thompson, S.S., & Walker-Dalhouse, D. (2011). Examining the trajectory of differentially skilled first graders' reading fluency of words in isolation and in context. *Reading & Writing Quarterly, 27,* 281-305. doi: 10.1080/10573569.2011.596095

McCarthy, D. (1972). *McCarthy Scales of Children's Abilities.* San Antonio, TX: The Psychological Corporation.

Mesmer, H.A.E. (2001). Decodable text: A review of what we know. *Reading Research and Instruction, 40*(2), 121-142.

Mesmer, H.A.E. (2005). Text decodability and the first-grade reader. *Reading and Writing Quarterly, 21,* 61-68. doi: 10.1080/10573560590523667

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups* (NIH Publication No. 00-4754). Retrieved from: http://www.nationalreadingpanel.org/Publications/subgroups.htm

Nese, J.F.T., Park, B.J., Alanzo, J., & Tindal, G. (2011). Applied curriculum-based measurement as a predictor of high-stakes assessment: Implications for researchers and teachers. *The Elementary School Journal, 111*(4), 608-624. doi: 10.1086/659034

Novak, H., Bonaventura, E., & Merenda, P.A. (1973). A scale for the early selection of children with learning problems. *Exceptional Children, 40,* 98-104.

Nurss, J.R. & McGauvran, M.E. (1986). *Metropolitan Readiness Tests.* San Antonio, TX: The Psychological Corporation.

Oakhill, J.V. & Cain, K. (2012). The precursors of reading ability in young readers: evidence from a four-year longitudinal study. *Scientific Studies of Reading, 16*(2), 91-121. doi: 10.1080/10888438.2010.529219

Pearce, L. & Gayle, R. (2008). Oral reading fluency as a predictor of reading comprehension on a state's measure of adequate yearly progress, *1*, 51-70.

Pearce, L.R., & Gayle, R. (2009). Oral reading fluency as a predictor of reading comprehension with American Indian and White elementary students. *School Psychology Review, 38*(3), 419-427.

Pearson. (2006, December 14). Harcourt Assessment acquires Edformation: With the addition of the AIMSweb product, Harcourt Assessment now provides a holistic assessment approach to response to intervention. Retrieved from http://www.pearsonassessments.com/haiweb/Cultures/en-US/Site/AboutUs/NewsReleases/NewsArchive/NewsRelease121406.htm

Pearson, P.D. (2006). Foreword. In K.S. Goodman (Ed.), *The truth about DIBELS: What it is, What it does.* (pp. v-xxiv). Portsmouth, NH: Henemann.

Pearson. (2012a). *AIMSweb technical manual.* Bloomington, MN: Author.

Pearson. (2012b). *AIMSweb reading curriculum-based measurement administration and scoring guide.* Bloomington, MN: Author.

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., & Feinstein, A.R. (1996). Simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*(12), 1373-1379.

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

Petscher, Y., & Kim, Y.S. (2011). The utility and accuracy of oral reading fluency score types in predicting reading comprehension. *Journal of School Psychology, 49,* 107-129. doi: 10.1016/j.jsp.2010.09.004

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, 36*(3), 158-166. doi:

10.1177/1534508410396698

Podhajski, B., Mather, N., Nathan, J., & Sammons, J. (2009). Professional development in scientifically based reading instruction: Teacher knowledge and reading outcomes. *Journal of Learning Disabilities, 42*(5), 403-417. doi:10.1177/0022219409338737

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, Second Edition.* Newbury Park, CA: Sage.

Raudenbush, S.W., Bryk, A.S, & Congdon, R. (2011). HLM 7 for Windows [Computer software]. Skokie, IL: Scientific Software International, Inc.

Reschly, A.L., Busch, T.W., Betts, J., Deno, S.L., & Long, J.D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47,* 427-469. doi: 10.1016/j.jsp.2009.07.001

Rice, J.K. (2010). *The impact of teacher experience: Examining the evidence and policy implications.* Retrieved from Urban Institute website: http://www.urban.org/uploadedpdf/ 1001455-impact-teacher-experience.pdf

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

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,* 343-366.

Samuels, S. J. (2007). The DIBELS tests: Is speed of barking at print what we mean by reading fluency? *Reading Research Quarterly, 42*, 563-567.

Samuels, S.J. (2004). Toward a theory of automatic information processing in reading, revisited.

In R.B. Ruddell & N.J. Unrau (Eds.), *Theoretical models and processes of reading, 5th Ed* (pp. 1127-1148). Newark, DE: International Reading Association.

Schrank, F. A., Mather, N., & Woodcock, R. W. (2004). *Woodcock- Johnson III Diagnostic Reading Battery: Comprehensive manual.* Itasca, IL: Riverside.

Shanahan, T., Callison, K., Carriere, C., Duke, N. K., Pearson, P. D., Schatschneider, C., & Torgesen, J. (2010). *Improving reading comprehension in kindergarten through 3rd grade: A practice guide* (NCEE 2010-4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from whatworks.ed.gov/publications/practiceguides.

Shapiro, E.S. (2011). *Academic Skills Problems: Direct Assessment & Intervention (4th ed.)*. New York: Guilford Press.

Shapiro, E.S., Solari, E., & Petscher, Y. (2008). Use of a measure of reading comprehension to enhance prediction on the state high stakes assessment. *Learning and Individual Differences, 18*, 316-328. doi: 10.1016/j.lindif.2008.03.002

Shapiro, E.S., Keller, M.A., Lutz, J.G., Santoro, L.E., & Hintze, J.M. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment, 24*(1), 19-35. doi: 10.1177/0734282905285237

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

Shinn, M.R. (2009, February). *Symposium on research practices and improvements in early literacy assessment.* Symposium presented at the National Association of School Psychologists 2009 Annual Convention, Boston, MA.

Shinn, M.R. (2012, March). *Use of highly decodable reading passages: Authentic assessment of early reading*. Presentation for the Center for Promoting Research to Practice Colloquium at Lehigh University, Bethlehem, PA.

Snow, C.E., Burns, M.S., & Griffin, P. (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Spear-Swerling, L. (2009). A literacy tutoring experience for prospective special educators and struggling second graders. *Journal of Learning Disabilities, 42*(5), 431-443. doi: 10.1177/0022219409338738

Spear-Swerling, L. & Sternberg, R.J. (1994). The road not taken: An integrative theoretical model of reading disability. *Journal of Learning Disabilities, 27*(2), 91-103. doi: 10.1177/002221949402700204

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), 585-607. doi: 10.1086/659032

Stevens, J.P. (2009). *Applied multivariate statistics for the social sciences, 5th Ed.* New York: Routledge.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47,* 522–532.

Torgesen, J.K. (1998). Catch them before they fall: Identification and assessment to prevent reading failure in young children. *American Educator/American Federation of Teachers*, Spring/Summer, 1-8.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of word reading efficiency.* Austin, TX: Pro-Ed.

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.

VanDerHeyden, A.M. (2010). Determining early mathematical risk: Ideas for extending the research. *School Psychology Review, 39*(2), 196-202.

Wang, Z. & Thompson, B. (2007). Is the Pearson r^2 biased, and if so, what is the best correction formula? *The Journal of Experimental Education, 75*(2), 109-125

Waterman, B.B. (2012). Review of the Group Reading Assessment and Diagnostic Evaluation. In *The Fifteenth Mental Measurements Yearbook.* Available from http://search.ebscohost.com

Williams, K. (2001). *Group Reading Assessment and Diagnostic Evaluation.* Circle Pines, MN: American Guidance Service.

Woodcock, R. W. (1973). *Woodcock Reading Mastery Tests*. Circle Pines, MN: American Guidance Service.

Woodcock, R. W. (l987). *Woodcock Reading Mastery Tests* (Rev. ed.). Circle Pines, MN: American Guidance Service.

Appendix A

HD Passages

Ann's mom got a rug. Ann sat on the rug. The 11
rug felt soft. Ann sat by a bug. Nat the bug had on 24
a hat. 26

Ann saw six bugs. The bugs had on hats. The 36
bugs had on red and black hats. 43

"I am mad," mom yelled. "I do not want bugs 53
on my rug." 56

Ann's mom picked up the rug. Mom hit the rug. 66
Mom hit the rug a lot. The bugs hung on the rug. 78

"We will not jump off!" the bugs yelled. 86

"I will wash the rug," said Ann's mom. "The 95
bugs must go." 98

Mom ran water. The rug got wet. The bugs got 108
wet. 109

Ann's mom left the rug in the sun. The rug got 120
hot. "We like the sun," the bugs yelled. 128

Ann's mom put the rug back in the hall. Mom 138
got a cat. Mom told it to get the bugs. The cat sat 151
on the rug. The bugs hand cat a big hat. 161

My best pal is a dog. His name is Jet. I got 12
him at a farm. Jet is six. Jet is a big dog. He can 26
do a trick. If I snap, Jet will jump. 35

Jet is black with a spot. His spot is tan. It is on 48
his back. I know it is Jet when I see the tan spot. 61

When I go, Jet sits on the step. He wags his 72
long tail as soon as I come back. He runs and 83
jumps too. 85

I think Jet is the best pet. He is a good dog. 97
My Dad and I will teach him to sit next. Dad says, 109
"Jet will sit when you snap and yell sit." 118

Mom likes to walk Jet in the park. She walks 128
him at dusk. He is a nice pet for Mom, Dad, and 140
me. 141

I am Ted the man. I can tell Kim the rat is sad. 13
He is mad at me. I will make him see I am funny. 26
He will not be mad or sad. 33
Kim the rat will not go. I will let him eat and 45
dig. We can have fun in the sand. We can 55
make a lot of sand men. Kim will see how fun 66
I can be. Rats are not bad. Rats can be fun. 77
Rats are not pets. Dogs and cats are pets. 86
Rats play with rats, not kids. 92
It did not work. Kim is mad and sad. He 102
will pack and go home. Kim the rat will not be 113
back. He is not nice. 118
I will get a cat to chase away Kim. The cat will 130
run fast. Kim will be mad. The cat will not hurt 141
Kim. He will let him go. Bye Kim. 149

Table 1

Descriptive Statistics for Student-Administered Measures in the Winter and Spring

Measure	*N*	*M*	*SD*	Range	Skewness	Kurtosis
Winter HD Median	234	47.6	37.2	0-159.6	0.8	-0.3
Winter HD Accuracy	234	83.0%	21.0	0-100%	-2.2	5.3
Retest HD Median	96	58.3	41.9	0-158	0.5	-0.7
Retest HD Accuracy	96	85.6%	22.6	0-100%	-2.5	6.6
Spring HD Median	222	66.6	41.8	0-217	0.6	-0.3
Spring HD Accuracy	222	90.9%	12.9	0-100%	-2.8	12.2
Winter NWF CLS	216	56.9	35.8	4-143	1.0	0.1
Winter NWF WWR	216	15.3	13.9	0-50	1.0	0.0
Winter DORF WC	216	42.0	36.1	2-155	1.1	0.5
Winter DORF Accuracy	216	78.9	19.2	27-100%	-0.8	-0.3
Spring NWF CLS	215	74.6	37.2	11-143	0.5	-0.9
Spring NWF WWR	215	21.6	15.4	0-50	0.4	-1.0
Spring DORF WC	215	59.13	39.78	3-195	0.59	-0.36
Spring DORF Accuracy	215	86.6	16.1	21-100%	-1.5	1.7
GRADE Comprehension Standard Score	220	103.4	19.4	64-145	0.1	-0.9

Note: HD = Highly Decodable Passages; NWF = Nonsense Word Fluency; CLS = Correct Letter Sequences; WWR = Whole Words Read; DORF = DIBELS Oral Reading Fluency; GRADE = Group Reading Assessment and Diagnostic Evaluation

Table 2

Reliability Correlation Matrices Between HD Passages for Test-Retest and Alternate Form Reliability at Each Assessment Period

		Test-Retest Reliability		
		1	2	
1	Winter HD Median	--		
2	Retest HD Median	.98**	--	
		Winter Alternate Form Reliability		
		3	4	5
3	Winter HD Passage 1	--		
4	Winter HD Passage 2	.97**	--	
5	Winter HD Passage 3	.97**	.97**	--
		Retest Alternate Form Reliability		
		6	7	8
6	Retest HD Passage 1	--		
7	Retest HD Passage 2	.98**	--	
8	Retest HD Passage 3	.98**	.98**	--
		Spring Alternate Form Reliability		
		9	10	11
9	Spring HD Passage 1	--		
10	Spring HD Passage 2	.96**	--	
11	Spring HD Passage 3	.96**	.97**	--

Note: N = 94 to 232 depending on the correlation; HD = Highly Decodable Passages
**$p < 0.01$ (two-tailed)

Table 3

Convergent Validity Correlation Matrix

		1	2	3	4	5	6	7	8	9
1	Winter HD Median	--								
2	Winter DIBELS NWF CLS	.84**	--							
3	Winter DIBELS NWF WWR	.82**	.96**	--						
4	Winter DIBELS ORF	.96**	.83**	.81**	--					
5	Spring HD Median	.95**	.81**	.80**	.93**	--				
6	Spring DIBELS NWF CLS	.82**	.86**	.84**	.80**	.83**	--			
7	Spring DIBELS NWF WWR	.80**	.84**	.83**	.78**	.81**	.96**	--		
8	Spring DIBELS ORF	.93**	.79**	.78**	93**	.96**	.84**	.82**	--	
9	GRADE Comprehension Composite	.83**	.65**	.67**	78**	.84**	.71**	.73**	.84**	--

Note: N = 206 to 220 depending on the correlation; HD = Highly Decodable Passages; DIBELS = Dynamic Indicators of Basic Early Literacy Skills; NWF = Nonsense Word Fluency; CLS = Correct Letter Sounds; WWR = Whole Words Read; ORF = Oral Reading Fluency; GRADE = Group Reading Assessment and Diagnostic Evaluation

**$p < 0.01$ (two-tailed)

Table 4

Fixed Effects Estimates for HLM

Parameter	Estimate	*SE*	*t* value
For INTRCPT1, β_0			
INTRCPT2, γ_{00}	-8.1	71.0	-0.1
Percent Free/Reduced Lunch, γ_{01}	-0.1	0.1	-0.9
Years of Teaching Experience, γ_{02}	-0.2	0.3	-0.7
Teacher Gender, γ_{03}	3.6	11.4	0.3
Teacher Ethnicity, γ_{04}	4.2	10.0	0.4
For W_HD_MED slope, β_1			
INTRCPT2, γ_{10}	2.4	2.5	1.0
Percent Free/Reduced Lunch, γ_{11}	0.0	0.0	0.5
Years of Teaching Experience, γ_{12}	0.0	0.0	0.3
Teacher Gender, γ_{13}	0.0	0.1	0.2
Teacher Ethnicity, γ_{14}	-0.2	0.4	-0.6

N = 222. –2 log-likelihood = 1,782.72
**p* < .05; ***p* < .01

Table 5

Random Effects Estimates for HLM

Parameter	*SD*	Variance Component	χ^2
INTRCPT1, u_0	6.2	38.5*	20.8
W_HD_MED slope, *u1*	0.1	0.00	17.0
level-1, *r*	12.6	159.7	

Figure 1. Plot of the ROC curve for winter HD passage score to GRADE Comprehension Composite score at or above the 25th percentile versus below the 25th percentile

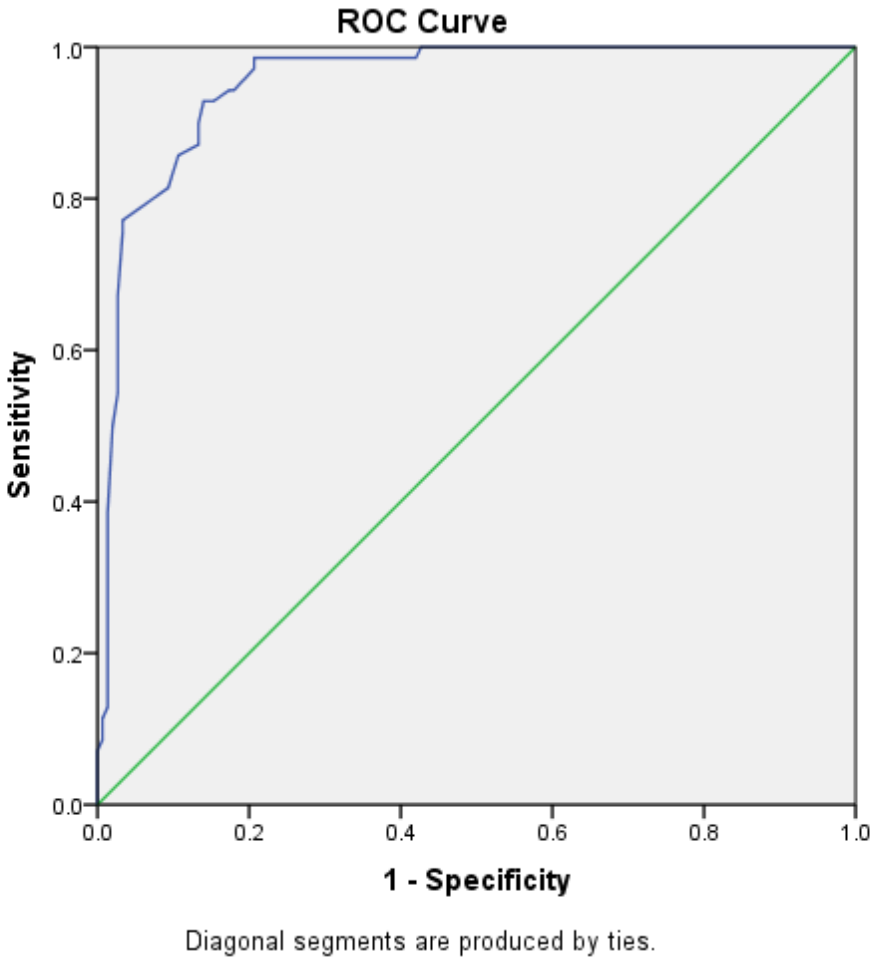
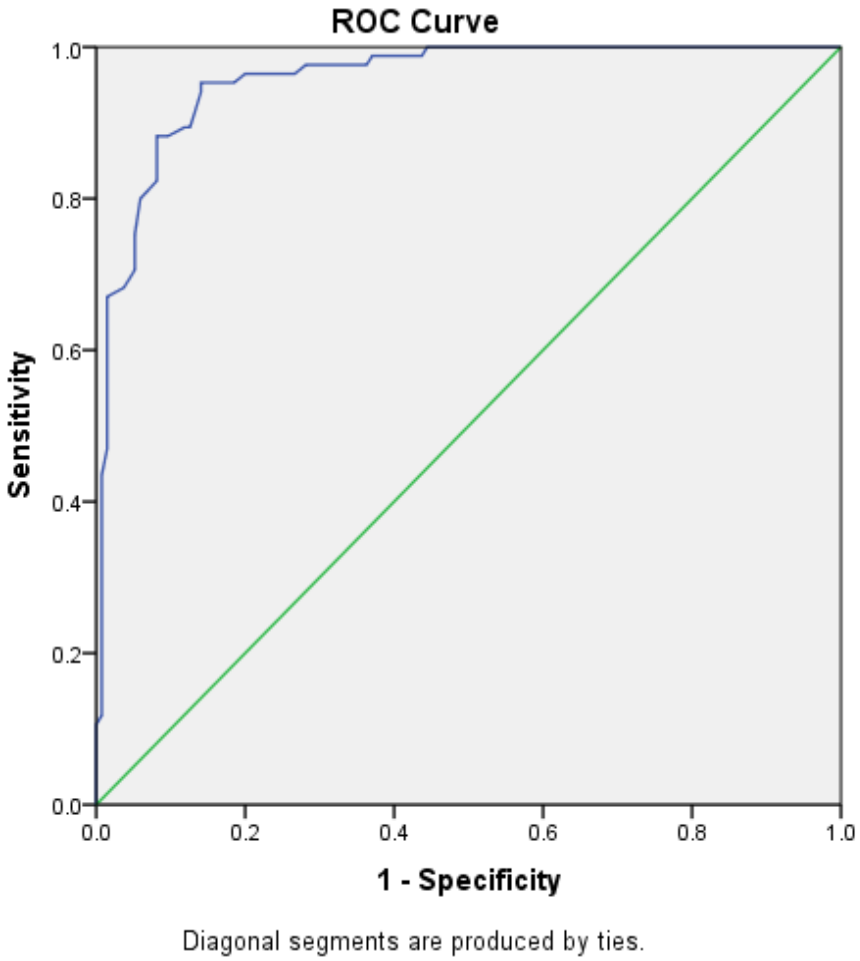


Figure 2. Plot of the ROC curve for winter HD passage score to GRADE Comprehension Composite score at or above the 40th percentile versus below the 40th percentile



Kirra B. Guard

70 Maxwell Avenue
Geneva, NY 14456

Kirra.guard@gmail.com
(315) 694-2781

EDUCATION

August 2009–Present — **Ph.D. Candidate, School Psychology**
Specialization in Students At-Risk or With Disabilities
Lehigh University, Bethlehem, Pennsylvania
APA accredited and NASP approved program
Expectated Graduation Date: September 2015

May 2009 — **M.A., Teaching**
Hobart and William Smith Colleges, Geneva, New York

May 2008 — **B.A., Summa Cum Laude, Psychology**
Hobart and William Smith Colleges, Geneva, New York
Phi Beta Kappa (2008)
Elizabeth Herendeen O'Dell Book Award (2008)
President's Public Service Award (2008)
Eta Sigma Phi (Classics Society) Member (2006)
Dean's List (Spring 2005-Spring 2008)
Hobart and William Smith Faculty Scholar
Admitted to William Smith Honor's House

PROFESSIONAL CERTIFICATES & SPECIALIZED TRAINING

July 2014 — **New York State School Psychologist, Provisional Certificate**
New York State Education Department, Office of Teaching Initiatives

May 2014 — **Childhood Education (Grades 1-6), Initial Teaching Certificate**
New York State Education Department, Office of Teaching Initiatives

May 2014 — **Students With Disabilities (Grades 1-6), Initial Teaching Certificate**
New York State Education Department, Office of Teaching Initiatives

September 2013 — **Pennsylvania School Psychologist Certification**
The Pennsylvania Department of Education

June 2012 — **Certified Positive Discipline Parenting Educator**
Positive Discipline Association

February 2011 — **Foundations of Disaster Mental Health Training**
American Red Cross

CLINICAL WORK EXPERIENCE

August 2013–June 2014 — **Graduate Assistant**
Centennial School of Lehigh University, Bethlehem, Pennsylvania
Site Supervisor: Julie Fogt, Ed.D

- Worked with faculty and staff to implement school-wide behavior support system

- Collaborated with teachers, students, and families to create plans for offering academic support to individual students
- Conducted comprehensive academic assessments of student skills for intervention planning
- Developed and implemented cognitive-behavior therapy sessions for groups of middle school students
- Provided professional development sessions on topics such as progress monitoring and linking assessment to intervention
- Offered individual support as needed to teachers on how to use AIMSweb for progress monitoring
- Consulted with individual teachers and school teams to develop individualized plans for supporting positive student behavior
- Analyzed school-wide academic progress monitoring data to evaluate strengths and needs of current educational services
- Collaborated with faculty and staff to write and present results of reevaluation reports

July 2012–July 2013 **Project Coordinator, Personnel Preparation in Response to Intervention**
Lehigh University, Bethlehem, Pennsylvania
Investigators: Edward Shapiro, Ph.D., Mary Beth Calhoon, Ph.D.

- Coordinated with school site supervisors to ensure understanding of requirements and approporiate graduate student opportunities
- Supervised specialist-level graduate students' practicum experiences
- Organized specialized training opportunities as necessary, such as DIBELS Next assessment training
- Supported and organized student efforts to produce a training video on RTII
- Monitored graduate student progress in terms of knowledge, skills, and level of comfort with RTII-related work
- Helped graduate students to develop and submit a poster proposal for 2013 NASP convention

January 2013–May 2014 **Teaching Assistant, SchP 425: Assessment and Intervention in Educational Consultation**
Lehigh University, Bethlehem, Pennsylvania
Instructor: Edward Shapiro, Ph.D.

- Communicated with cooperating schools to identify appropriate elementary school cases for graduate students enrolled in the course
- Supervised graduate students (Ed.S. and Ph.D.) in obtaining interviews, observational data, and assessment results
- Provided assistance in report writing, determining appropriate student interventions, and establishing a plan for progress monitoring
- Maintained communication with cooperating schools to ensure practicum cases are running smoothly, answer questions, and address concerns

July 2006–August 2006 **Assistant Preschool Teacher**
Geneva Recreation Center, Geneva, New York
Supervisor: Martha Wilson

- Supervised preschool children during play and structured learning activities
- Assisted lead teacher with organizing and implementing daily activities
- Communicated with families daily regarding child's activities and upcoming events and plans

CLINICAL/PROFESSIONAL TRAINING

July 2014–Present

Predoctoral Internship
LeMoyne Elementary School, Syracuse, New York
Site Supervisor: Kristi Cleary, Ph.D.

- Work with supervisor to develop school-wide behavioral management procedures
- Participate in school problem-solving and planning teams to develop school-wide interventions and individual student support plans
- Provide professional development seminars on positive behavior support and school-wide behavioral management procedures
- Consult with teachers to develop and implement classroom management plans
- Conduct functional behavioral assessments and consult with teachers to develop individual behavior interventions
- Conduct comprehensive psychoeducational evaluations to determine need for special education services
- Provide intensive individual academic interventions to students who require tier 3 services within a response to intervention system
- Develop and implement cognitive-behavior therapy sessions focused on anger management to elementary students in a self-contained classroom for students with emotional and behavioral difficulties
- Work individually with students to conduct problem solving, provide therapy, develop social skills, and offer other social-emotional interventions as needed

August 2012–June 2013

School Psychology Practicum
Centennial School of Lehigh University, Bethlehem, Pennsylvania
University Supervisor: Christine Novak, Ph.D.
Site Supervisor: Julie Fogt, Ed.D

- Conducted comprehensive academic assessments of student skills for intervention planning
- Consulted with teachers to develop and refine plans for effective behavioral interventions for individual students and assisted with progress monitoring
- Assisted with multidisciplinary reevaluations and individual education plan meetings for enrolled students
- Provided professional development to teachers on methods for linking assessment to instruction within an RTI framework, and on conducting behavioral observations in the classroom
- Consulted with administrators to develop a system for evaluating school-wide academic progress
- Supported senior staff in offering tours of Centennial School and consultative services regarding positive behavior support

August 2011–June 2013

School Psychology Practicum
Fountain Hill Elementary School, Bethlehem, Pennsylvania
University Supervisor: Christine Novak, Ph.D.
Site Supervisor: Michelle Lesinski, M.Ed.

- Conducted comprehensive, integrated student evaluations that included behavioral observation, interviews, curriculum-based assessment, rating scales, and standardized assessment
- Assisted with functional behavioral assessments, as well as intervention development and implementation
- Participated in multidisciplinary team evaluation and individual education plan meetings

- Provided consultation services to teachers, families, and administrators
- Provided RTI reading instruction in kindergarten for tiers 1, 2, and 3
- Lead year-long group and individual therapy sessions using CBT
- Assisted with crisis counseling in the district
- Conducted conflict management sessions with students
- Worked with local community agency to seek funding for programming for early childhood behavior problems based on teacher and parent concerns

January 2011–June 2011 **School Psychology Practicum in Behavioral Assessment**
Thomas Jefferson Elementary School, Bethlehem, PA
University Supervisor: Edward Shapiro, Ph.D.

August 2010–June 2011

School Psychology Practicum in Assessment and Intervention in Educational Consultation
Thomas Jefferson Elementary School, Bethlehem, PA
University Supervisor: Edward Shapiro, Ph.D.

- Conducted comprehensive evaluation that included interviews, observation, completion of rating scales, analogue assessment, direct assessment using curriculum-based measurement, and permanent product review
- Wrote and presented evaluation reports to family and school personel, including recommendations for intervention approaches
- Collaborated with teacher to implement a classroom-based intervention program for behavior
- Conducted weekly academic intervention sessions devoted to developing decoding skills
- Communicated outcomes of interventions to families and school through written intervention reports after several weeks of implementation and progress monitoring

August 2010–December 2010 **School Psychology Practicum in Consultation Procedures**
Head Start of Lehigh Valley, Bethlehem, PA
University Supervisor: Patricia Manz, Ph.D.

- Conducted problem identification and conjoint needs identification interviews
- Conducted problem analysis and conjoint needs analysis interviews
- Collaborated with school personnel and family to develop and implement a school-based intervention, which was extended to the student's home
- Collected data to monitor progress and evaluate acceptability of the intervention and consultation process

January 2010–May 2010 **School Psychology Practicum in Assessment of Intelligence**
University Supervisor: Kevin Kelly, Ph.D.

- Conducted full battery assessments for four individuals, including an elementary school student, middle school student, high school student, and an adult
- Submitted written reports and completed protocols for feedback and revision to improve clinical assessment skills and writing abilities

November 2008–January 2009 **Student Teacher, Fifth Grade Special Education Classroom**
West Street Elementary School, Geneva, New York
Univeristy Supervisor: Mary Kelly, Ph.D.
Site Supervisor: Margaret Jarecke, M.Ed.

- Maintained daily schedule, including pulling students from classes for small group instruction and returning them to class on time
- Managed group and individual behavior plans
- Created daily lesson plans aligned with student IEPs, state standards, and

district scope and sequence
- Coordinated with supervisor to manage the development of new IEPs, as well as changes to existing IEPs
- Collaborated with general education teacher to develop lesson plans appropriate for both general education and special education students as a whole group
- Managed communication with parents regarding student performance
- Attended district meetings for student instructional support, special education referral, and individual education plan development

August 2008–October 2008 **Student Teacher, Fifth Grade General Education Classroom**
West Street Elementary School, Geneva, New York
Univeristy Supervisor: Mary Kelly, Ph.D.
Site Supervisor: Marlene Young, M.Ed.
- Maintained responsibility for daily scheduling, including managing time for lessons and ensuring students arrived at specials and lunch on time
- Developed and implemented classwide behavior management system
- Created daily lesson plans aligned with state standards and district scope and sequence
- Managed communication with parents regarding student performance
- Attended district meetings for curricula development, student instructional support, and special education referral

RESEARCH EXPERIENCE

January 2013–Present **Dissertation**
Lehigh University, Bethlehem, Pennsylvania
Advisor: Edward Shapiro, Ph.D.
Title: Highly Decodable Reading Passages as a First-Grade Screening Measure: A Validation Study
- Conducted comprehensive review of research on reading screening measures for early elementary school students
- Worked with advisor and committee members to develop the methods for evaluating psychometric properties of the measures
- Collaborated with statistician, advisor, and committee members to establish a plan for analyzing the data
- Recruited area elementary schools to participate in the study

January 2010–August 2013 **Research Assistant, Rating Scales for Academic Skills**
Center for Promoting Research to Practice
Lehigh University, Bethlehem, Pennsylvania
Investigators: Edward Shapiro, Ph.D., Sandra Chafouleas, Ph.D., Chris Riley-Tillman, Ph.D.
- Reviewed research on mathematics skill development and instruction in the elementary grades
- Researched various state and national mathematics standards for grades 3 through 5
- Developed initial math and reading rating scale items and related domains
- Completed domain and item validation with content experts and area teachers
- Refined domain and item wording based on feedback and research findings
- Worked with fellow students and lead researchers to present results of work at national conventions
- Conducted pilot testing to evaluate the math rating scale for students in

grades 3 through 5
- Analyzed data to examine predictive validity, content validity, and reliability of math rating scale

August 2009–August 2013 **Research Assistant, Project READERS**
Center for Promoting Research to Practice
Lehigh University, Bethlehem, Pennsylvania
Investigators: Edward Shapiro, Ph.D., Todd Glover, Ph.D., Tanya Ihlo, Ph.D., and Stacy Martin, Ph.D.
- Participated in meetings on initial study design and methodology
- Assisted lead researchers in reviewing and compiling relevant research
- Helped to develop and refine survey items on teacher knowledge and logs of classroom practices
- Developed and submitted IRB proposals for review
- Collaborated with lead investigators and project coordinators to recruit schools to participate in the study
- Worked with the project coordinator to organize and implement teacher training sessions held at Lehigh University for participating teachers
- Conducted standardized reading assessments with participating students

Research Assistant, Teachers SPEAK
August 2009–August 2013 ***Center for Promoting Research to Practice***
Lehigh University, Bethlehem, Pennsylvania
Investigators: Edward Shapiro, Ph.D., Todd Glover, Ph.D., Tanya Ihlo, Ph.D., and Guy Trainin, Ph.D.
- Participated in meetings on initial study design and methodology, as well as follow-up meetings to review and revise initial plans
- Assisted lead researchers in reviewing and compiling relevant research
- Helped to develop and refine survey items on teacher professional development experiences, knowledge, and classroom practices
- Developed and submitted IRB proposals for review
- Participated in meetings reviewing results of survey data
- Contributed feedback to working article drafts reporting results of the national survey study

August 2008–May 2009 **Masters Thesis**
Hobart and William Smith Colleges, Geneva, New York
Advisor: Julie Newman Kingery, Ph.D.
Thesis: Animal Assisted Activity in the Classroom: Effects on School Involvement, Motivation, and Emotional-Behavioral Adjustment
- Developed a pretest-posttest study design on animal assisted therapy/animal assisted activity in the classroom for at-risk youth
- Obtained consent from thirty at-risk middle school students and their guardians for students to participate in the study
- Identified assessment measures and consulted with school officials to design project
- Collaborated with teachers to collect survey data and introduce certified therapy dog
- Utilized SPSS for all data entry and analyses

May 2007–July 2008 **Research Assistant, Centennial Center for Leadership Research**
Hobart and William Smith Colleges, Geneva, New York
Research Advisor: Robert Murphy, M.Ed.
- Conducted background research on over 25 college and university leadership centers
- Reviewed NASPA and Kellogg Foundation research on leadership development programs and developed interview questions based on research findings

- Interviewed 18 leadership development program directors and members of the Hobart and William Smith Colleges' community
- Presented research results to Colleges' senior staff, including president, provost, and deans

GRANTS & SCHOLARSHIPS

Pearson Clinical Assessments Annual Trainers of School Psychologists Professional Development Scholarship (received January 25, 2015), Lead Presenter, *Highly decodable passages: A new screening measure for early readers*, NASP 2015 Annual Convention, $500.

Society for the Study of School Psychology Dissertation Grant Award (received November 26, 2013), Co-Investigator, Highly Decodable Reading Passages as a First-Grade Screening Measure: A Validation Study, $4,000.

PUBLICATIONS

Shapiro, E.S., Gebhardt, S.N., **Guard, K.B.,** Flatley, K., Fu, J., & Leichman, E.S. (in preparation). *Development and validity of the Rating Scales of Academic Skills: Two pilot investigations.*

Shapiro, E.S. & **Guard, K.B.** (2014). Best practices in setting progress monitoring goals for academic skill improvement. In A. Thomas & P. Harrison (eds.), *Best Practices in School Psychology VI.* Washington, DC: National Association of School Psychologists.

CONFERENCE PRESENTATIONS

Guard, K.B., & Shinn, M.R. (February 2015). *Highly decodable passages: A new screening measure for early readers*. Paper session presented at the National Association of School Psychologists 2015 Annual Convention, Orlando, FL.

Fogt, J. B. & **Guard, K.B.** (February 2015). *Reading gains of students with EBD within a SWPBS Framework*. Paper session presented at the National Association of School Psychologists 2015 Annual Convention, Orlando, FL.

Shapiro, E.S., & **Guard, K.B.** (February 2014). *Technological advancements in conducting direct systematic observations*. Mini skills session presented at the National Association of School Psychologists 2014 Annual Convention, Washington, DC.

Shapiro, E.S., & **Guard, K.B.** Contributors: Calhoon, M.B., & Leichman, E. (February 2014). *Utilizing teacher judgment: The Rating Scales of Academic Skills*. Paper session presented at the National Association of School Psychologists 2014 Annual Convention, Washington, DC.

Van Oss, E., Hunter, B., Nuschke, A., & **Guard, K.B.** (February 2014). *SLD identification: Case studies comparing RTI and traditional data*. Poster session presented at the National Association of School Psychologists 2014 Annual Convention, Washington, DC.

Shapiro, E.S., **Guard, K.B.,** Gebhardt, S., & Leichman, E. (July 2013). *Development of rating scales for academic assessment: Reading comprehension and pre-algebra skills*. Poster session presented at the 2013 annual meeting of the American Psychological Association, Honolulu, Hawaii.

Guard, K.B., Hermetet-Lindsay, K.D., & Krehbiel, C.F. (2013, February). *Academic skills and mental health promotion through critical media literacy*. Paper session presented at the National Association of School Psychologists 2013 Annual Convention, Seattle, WA.

Guard, K.B., Leichman, E.S., Shapiro, E.S. (2013, February). *Initial development of a brief rating scale for mathematics skills***.** Poster session presented at the National Association of School Psychologists 2013 Annual Convention, Seattle, WA.

Guard, K.B., Krehbiel, C.F., & Hermetet-Lindsay, K.D. (2012, February). *Media literacy in adolescence: A review of the literature*. Poster session presented at the National Association of School Psychologists 2012 Annual Convention, Philadelphia, PA.

PROFESSIONAL AFFILIATIONS AND LEADERSHIP ROLES

January 2013–March 2015
Student Review Board Member
Assessment for Effective Intervention

August 2011–May 2012
Student Representative, APA Division 16 (School Psychology)
Lehigh University, Bethelehem, Pennsylvania
- Maintained communication with APA Division 16 organizers and members
- Relayed important information and opportunities to students and faculty within the Lehigh University School Psychology program

August 2011–July 2012
Vice President, Lehigh University Student Affliates in School Psychology
Lehigh University, Bethelehem, Pennsylvania

August 2010–July 2011
President, Lehigh University Student Affliates in School Psychology
Lehigh University, Bethelehem, Pennsylvania
- Coordinated with APA, Division 16 to maintain Lehigh University's SASP chapter
- Collaborated with other officers and members to organize and schedule events for the year
- Attended meetings held by the University Graduate Life Office to stay up to date on club requirements and to communicate information regarding events
- Planned, organized, and publicized academic activities and professional development events related to school psychology
- Organized community service events
- Sponsored and participated in social gatherings intended to foster positive relationships among graduate students

2010–Present
Member, American Psychological Association

Member, Division 16 of the American Psychological Association

Member, Association of School Psychologists of Pennsylvania

Member, National Association of School Psychologists

Member, Pennsylvania Psychological Association

PROFESSIONAL REFERENCES

Edward Shapiro, Ph.D.
Professor & Director
Center for Promoting Research to Practice
School Psychology Program
College of Education
Lehigh University
Iacocca Hall, Room L-111
111 Research Drive
Bethlehem, PA 18015
(610) 758-3258
ed.shapiro@lehigh.edu

Kristi Cleary, Ph.D.
School Psychologist
LeMoyne Elementary School
1528 LeMoyne Ave.
Syracuse, NY 13208
(315) 435-4980
kcleary@scsd.us

Julie Fogt, Ed.D.
School Psychologist
Centennial School of Lehigh University
2196 Avenue C, LVIP I
Bethlehem, PA 18017
(610) 266-6500
juf2@lehigh.edu

Michelle Lesinski, M.Ed.
School Psychologist
Fountain Hill Elementary School
Bethlehem Area School District
1330 Church Street
Bethlehem, PA 18015
(610) 865-5881
mlesinski@beth.k12.pa.us