UNIVERSITY OF CALIFORNIA
RIVERSIDE

Early Response-to-Intervention Measures and Criteria as Predictors of Reading Disability in 3rd Grade

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Education

by

Kristen Dawn Beach

June 2012

Dissertation Committee:
Dr. Rollanda O'Connor, Chairperson
Dr. H. Lee Swanson
Dr. George Marcoulides

The Dissertation of Kristen Dawn Beach is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgements

I would like to acknowledge my advisor, Dr. Rollanda O'Connor, for her unwavering support and guidance throughout my career as a doctoral student. I would also like to thank her for holding the highest of expectations of my work and of me, thus contributing to my academic, professional, and personal growth over the past 3 years.

I would also like to acknowledge the tremendous faculty in the Graduate School of Education at the University of California, Riverside. Their expertise, support, and patience create an academic environment in which every student can excel.

A special thank-you to my peers, who reviewed numerous drafts of work and shared countless stories of failure and success, all of which contributed to my growth and sanity as a doctoral student.

Finally, I would like to acknowledge the research staff who worked the Response-to-Intervention project on which this dissertation is based. Thank you to Dr. Kathleen Bocian, who was and is a cherished leader and friend. Thank you to the school leaders, teachers, and students, who fed my passion for work in special education, and without whom this dissertation would not exist.

Dedication

I dedicate this dissertation to my past students, who inspired me to pursue a doctoral degree so that I could influence the educational system that held limited expectations for their academic performance. For the students who I tutored during my work at UCR, all of whom contributed to my understanding of student learning and behavior. For my mother, who taught me perseverance and acceptance. And for my father, who contributed to my personal strength and determination, and who was always my number one fan.

ABSTRACT OF THE DISSERTATION

Early Response-to-Intervention Measures and Criteria
as Predictors of Reading Disability in 3rd Grade

by

Kristen Dawn Beach

Doctor of Philosophy, Graduate Program in Education
University of California, Riverside, June 2012
Dr. Rollanda O'Connor, Chairperson

Reading is the most valuable skill children must master early in schooling. Unfortunately, many students struggle to read and may be identified as having a Reading Disability (RD). In this dissertation, I explored the usefulness of the Response-to-Intervention (RtI) framework for identifying children with RD by examining the use of 1st and 2nd grade reading measures and responsiveness criteria for predicting RD in 3rd grade. Data were derived from a longitudinal RtI project executed in low-income, high-poverty schools in Southern California. Participants attended one of five schools from 1st to 3rd grade and had access to a high-quality Tier II intervention during their attendance. I used logistic regression to identify reading measures most useful for predicting RD in word reading/fluency (WR-F) and comprehension/vocabulary (C-V) separately; I then paired intervention responsiveness criteria with significant predictors and explored RD classification accuracy using 2x2 contingency tables. Model-based results generally yielded superior classification accuracy compared to single-measure predictors of RD;

however, 1st grade word identification and 2nd grade oral reading fluency showed promise as isolated measures for predicting RD in WR-F. Model-based predictions were required to obtain adequate classification accuracy for RD in C-V. While the former finding is promising for early identification of those students in need of more intensive instruction in lexical or fluency-based skills, the latter finding reaffirms literature attesting to the complexity of RD in comprehension and difficulty of predicting such deficits using early measures of reading, which principally assess word reading skill. Models and classification analyses replicated well with an independent sample, thus enhancing confidence in study results. Practical implications and need for future research are discussed.

Table of Contents

Chapter 1: Introduction

Learning to read is one of the most essential activities that young children are expected to master early in schooling. Reading sets the foundation for acquiring knowledge in content areas in late elementary school (Chall, 1996) and is required to become a fully functioning member of society. Unfortunately, there are some children who will struggle to read even after receipt of good instruction. When a child exhibits an extreme academic deficit that is unexpected based on his potential and/or not easily remediated through intervention, he may be eligible for special education services under the eligibility category of Learning Disability (LD).

The Individuals with Disabilities Education Improvement Act (IDEIA, 2004) defines LD as "a disorder in one or more of the basic psychological processes involved in understanding or in using language, spoken or written, which may manifest itself in the imperfect ability to listen, think, speak, read, write, spell, or do mathematical calculations." The definition also stipulates that the disorder cannot be secondary to another disability or to economic hardship [IDEIA, 2004, sec 300.8(c)(10)]. Although children may struggle in a number of areas, most children (up to 90%; Lerner, 1989) classified as LD struggle in the area of reading. For the purpose of this dissertation, LD and RD will be used interchangeably.

Historically, children presumed to have an LD became involved in a problem solving process that included identification of the area of need, provision of academic and other supports, and IQ and academic testing to determine whether an LD existed (Ikeda, Rahn-Blaskeslee, Neibling, Gustafson, Allison, & Stumme, 2007). As of the

reauthorization of IDEIA in 2004, states were granted the right to pursue alternative means for identifying children with LD: Response to Intervention (RtI). The purpose of this dissertation is to explore the validity of the RtI framework for identifying children with an LD, specifically with LD in reading (i.e. RD). Although accepted by many researchers as a preventative framework, RtI has yet to demonstrate functionality as a mechanism for identifying RD. Using longitudinal data collected from two cohorts of children in grades 1-3, I explored combinations of RtI measures and criteria to identify those that best predicted the long-term reading skill of students who had access to reading intervention in 1st through 3rd grades. I replicated results on a second cohort of students attending the same schools and receiving the same reading intervention to test the reliability of the results. Research questions were as follows:

1) What combination of reading assessment and criteria collected from students receiving access to reading intervention from 1st to 3rd grade demonstrates the most adequate sensitivity and specificity rates for predicting those children who will be identified as proficient or RD readers at the end of grade 3?
2) Do these findings replicate on a different cohort of students receiving the same intervention in the same schools during approximately the same time period (1 year delay)?

I begin this dissertation by describing how LD was historically identified and how the classification of students as LD has undergone change in recent years. I then describe the RtI framework and explain how it might be used as a vehicle for RD identification. Finally, I review the literature that describes issues with using the RtI framework as an

identification tool. This review is followed by the current study's research method, results, and discussion.

History of LD Identification

Before the reauthorization of the Individuals with Disabilities Education Improvement Act (IDEIA, 2004; P.L. 108-446) children who struggled academically in school were almost exclusively referred to special education after 1) being identified as struggling by their teacher of record, 2) "reasonable efforts" had been made to improve the child's academic performance in the general education environment, and 3) meeting eligibility criteria. Eligibility was established after confirming a marked difference (typically 1.5 to 2 standard deviations) between the child's ability, measured by IQ tests, and his/her academic achievement as measured by standardized, normed achievement tests.

Although widely used as special education eligibility formula, the IQ-Discrepancy (IQ-D) model for identifying children in need of specialized instruction has met considerable criticism in the past twenty years. First, researchers have noted an over-representation of ethnic minorities and children living in low socioeconomic (SES) environments when IQ-D is used to refer students to special education (Donovan & Cross, 2002; Speece & Case, 2001). Also, the IQ-D model has been criticized as a "wait-to-fail" model (Fuchs & Fuchs, 2006; Gresham, 2002) – often, children must fail to make gains in general education for several years before meeting the IQ-D criterion and becoming eligible for services. Some researchers challenge the assumption of linearity between IQ and reading ability – that is, the reading ability of children with below

average, average, and high IQ does not follow a similar below average, average, and above average pattern (see Vellutino et al., 1996). A similar finding is that many poor readers have comparable early reading profiles, yet differ in their satisfaction of IQ-D criteria (Fletcher et al., 1994; O'Malley, Francis, Foorman, Fletcher, & Swank, 2002). Overall, early measures of reading skill (e.g. phonemic awareness) more consistently identify struggling readers compared to IQ-D models (Fletcher et al., 1994; Vellutino et al., 1996). These findings challenge the logic of using the IQ-D model to diagnose RD. Gresham (2002) also noted that the IQ assessments required for eligibility tell educators very little about how to instruct struggling students; he advocated for an identification approach that links eligibility assessment results and future instructional interventions. Despite these issues, many states rely on the IQ-D formula to identify children who may need special education services and students who satisfy IQ-D may become eligible for services under the LD eligibility category.

For many years and for many of the reasons stated above, researchers have been exploring alternative strategies for identifying children as "at-risk" or in need of special education services. With the reauthorization of IDEIA in 2004, the federal government condoned the use of an alternative strategy for identifying children in need of special services. The strategy that is being adopted faster than a strong research base for its effectiveness as a special education referral tool can be compiled is RtI.

Response to Intervention

The RtI framework has been the leading contender for the replacement of the IQ-D method for identifying students with LD. Though also conceptualized as a mechanism

for providing high-quality early intervention, the federal government has officially approved states' use of the RtI model as a method of disability identification (IDEIA, 2004). The RtI framework relies on data-driven decision making and subsequent provision of increasingly intense instructional environments ("layers" of interventions; O'Connor, 2000) to best meet children's instructional needs. RtI typically comprises three layers (or tiers) of increasingly intensive and individualized instruction (Fuchs & Fuchs, 2006; O'Connor, Fulmer, Harty & Bell, 2005; Vaughn, Wanzek, Woodruff, & Linan-Thompson, 2007). Tier I is most commonly associated with the general education environment. In this tier, general educators provide presumably high-quality instruction to all students within the general education setting. In some cases, whole-class general education interventions are delivered to enhance the quality of Tier I instruction (see Mathes, Torgesen & Allor, 2001; McMaster, D. Fuchs, L. Fuchs, & Compton, 2005). Additionally, general educators may be provided with extensive professional development to improve the quality of Tier I instruction and/or collection and use of progress monitoring data for all students (see Mathes, Denton, Fletcher, Anthony, Francis, & Schatschneider, 2005; O'Connor, 2000; O'Connor et al., 2005; Vaughn et al., 2007). In many cases, increasing the quality of the general education instructional environment has led to student gains in reading (McMaster et al., 2005; O'Connor, 2000; O'Connor et al., 2005), with larger effects on student reading growth when teachers are involved in multiple years of professional development (Vaughn et al., 2008). Lastly, Tier I is the home of universal screening, where all students are tested on measures of early reading skill (e.g. letter-word knowledge) from kindergarten and consistently

throughout each year at each grade to monitor the effect of Tier I instruction as well as to screen for students who are not making adequate progress.

Even when improvements are made to Tier I instruction, there are often students who continue to struggle. Students who fail to make adequate progress (compared to peers at local, state, or national level norms and as measured by the Tier I universal screening) are identified as needing more intensive instruction, which is then offered at Tier II. Secondary intervention (i.e. Tier II in an RtI model) has been executed in at least two ways. Before RtI, educators implemented an individualized intervention designed (perhaps by a multidisciplinary team, such as a Student Support Team following a problem solving approach) to meet the precise needs of a single student (Fuchs, Mock, Morgan, & Young, 2003; Ikeda et al., 2007). Under RtI, a standardized treatment protocol is administered in a small group or one-on-one format (Al Otaiba & Fuchs, 2006; Fuchs et al., 2003; Vellutino et al., 1996). Wanzek and Vaughn (2007) described a standardized treatment as one that specifies "a-priori, the elements of reading instruction that will be implemented" (p. 542). The curriculum is research-based and fidelity of implementation is recorded. Sometimes researchers implement interventions that are more responsive to students' needs and that are less standardized (Mathes et al., 2005; O'Connor et al., 2005; O'Connor, Sanchez, & Bocian, submitted; Vaughn, Linan-Thompson, Kouzekanani, Bryant, Dickson, & Blozis, 2003), but do not fall under the individualized intervention title. These interventions, better classified as "low standardization" (Wanzek & Vaughn, 2007), have similar effect sizes as highly standardized treatments (Mathes et al., 2005; Wanzek & Vaughn, 2007). In general,

proponents of the standardized treatment protocol (e.g. Fuchs & Fuchs, 2006; Fuchs et al., 2003) suggest that this method, compared to the individualized problem-solving approach, may be more feasible for general educators and other school professionals to perform since teachers have to learn and implement only one or a few standardized protocols rather than being required to develop individualized interventions for every struggling student.

Tier II instruction is more intensive than Tier I instruction in several ways. First, instruction is more teacher-directed, systematic, and explicit (Fuchs & Fuchs, 2006). Next, instruction is provided in small group or one-to-one settings, which according to several studies leads to stronger student gains than the same instruction provided in larger groups for struggling students (Juel & Minden-Cupp, 2000; Vaughn, Linan-Thompson, Kouzekanani, et al., 2003) and for students with diagnosed disabilities (Elbaum, Vaughn, Hughes, & Moody, 1999; Thurlow, Ysseldyke, Wotruba, & Algozzine, 1993). Additionally, instruction may occur more frequently, may last longer per session and per intervention period, and/or may target a struggling student's specific area of need. Typically, reading interventions that occur frequently and last for several months have demonstrated positive effects (O'Connor, 2000; Vaugh, Linan-Thompson, & Hickman, 2003; Wanzek & Vaughn, 2007), though not all students respond to these well orchestrated interventions. Wanzek and Vaughn (2008) note that simply increasing intervention time without a change in instruction or group size may do little to promote the reading skills of more severely impaired readers. For these readers, Torgesen, Alexander, Wagner, Rashotte, Voeller, & Conway (2001) found that an intense and

individualized intervention administered one-to-one and totaling 67.5 hours significantly improved reading scores.

Any student who makes inadequate progress after receiving the more intensive, high-quality instruction at Tier II may then receive Tier III services. In some conceptualizations of RtI, Tier III is deemed "special education" and the student might be referred to a multidisciplinary team for additional assessment and referral for special education (Fuchs & Fuchs, 2006). In other cases, Tier III is simply additional intervention targeting students' reading needs (Denton, Fletcher, Anthony, & Francis, 2006), often offered in a one-to-one or small group setting (Kamps, Abbott, Greenwood, Wills, Veerkamp, & Kaufman, 2008). Currently, students are only identified as needing Tier III services if they have failed to thrive in high-quality Tier I and Tier II educational environments. Some researchers (e.g. Compton et al., submitted) propose that students may be "fast-tracked" to Tier III based on their academic profile in response to Tier I instruction and progress monitoring. In general, however, struggling students move through Tiers I and II before becoming eligible for Tier III services. Because Tier III is often the level at which students are considered for special education eligibility, research must determine whether students with LD can reliably be identified via failure to thrive *before* reaching Tier III – namely, in the Tier I and II environments. Therefore, I will focus on students' responsiveness-to-intervention provided in those settings.

Like IQ-Discrepancy, the RtI framework has met many criticisms, especially regarding its use as a tool to identify students with LD. There are concerns regarding whether measures and criteria for judging RtI and classifying children as RD are valid

and reliable. In current research, groups of good and poor responders to Tier I (Brown-Waesche, Schatschneider, Maner, Ahmed, & Wagner, 2011) and Tier II (Barth, Stuebing, Anthony, Denton, Mathes, Fletcher, & Francis, 2008; D. Fuchs, L.S. Fuchs, & Compton, 2004; Fuchs, Compton, Fuchs, Bryant & Davis, 2008) interventions vary depending on the measures and criteria researchers use to designate these groups. Furthermore, researchers have noted instability in group membership as students advance through grades, at least when response to Tier I instruction is explored (Brown-Waesche et al., 2011; Francis, Fletcher, Stuebing, Lyon, B.A. Shaywitz, & S.E. Shaywitz, 2005). If RtI is to be used as a mechanism for identifying RD, there should be some consistency with which students are identified. This consistency might be achieved through use of common measures and criteria for assessing students' responsiveness to intervention, at least within certain populations. Below I outline research investigating the usefulness of RtI as a LD identification tool with a focus on the impact of measures, criteria, and cut-offs for forming groups of good and poor responders and for describing overall reading proficiency.

Chapter 2: Review of the Literature

Within the RtI framework, several considerations must be made before deciding that instruction delivered in Tiers I and II is insufficient to meet a struggling student's needs. These considerations include: 1) the nature, intensity, and quality of the instruction/intervention; 2) the assessments used to measure progress and performance; and 3) the criteria applied to each assessment to designate students as good or poor responders. Criteria to designate students as good and poor responders varies with regard

to the reference group to which the struggling student is compared and the timing of the assessment process. Measures administered post-intervention are often used to establish the student's level of reading proficiency.

Nature and Intensity and Quality of Tier I and II Instruction/Intervention

RtI models assume students are first exposed to high-quality instructional environments. Therefore, if students are to be judged as "failing to respond" to general education instruction, the quality of the instruction must be established. The average rate of growth on some assessment should be tracked to establish whether most students are making adequate growth in the general education classroom; classrooms that have comparatively poor mean rates of growth compared to other classes in the school, district, state, or nation may need Tier I intervention first to develop a stronger instructional program for all students (Fuchs, Fuchs, & Compton, 2004). Only then is there confidence in the referral of students who do not benefit from Tier I instruction to receive Tier II intervention services. Although the quality of the general education context is assumed in RtI models, it is rarely assessed or reported in research or practice (Speece, Pericola-Case, & Molloy, 2003). In the instances where Tier I instruction was monitored, its relationship with Tier II eligibility has been unstable. For example, Al Otaiba and Fuchs (2006) measured the quality of Peer Assisted Learning Strategies (PALS) instruction, a general education intervention designed to improve the reading skills of kindergarten and 1st grade children, as implemented by the general education teacher (i.e. Tier I). They found that poor responders were most often in classrooms with the teachers who implemented the intervention with poor quality. Similarly in their

ethnography detailing the special education referral process, Harry and Klingner (2006) reported that special education referrals most often came from teachers with poor instructional ability and inadequate classroom management. Therefore, the quality of the Tier I instructional environment appears linked to a child's probability of being referred for intensive intervention under the RtI model, or special education under alternative referral models.

Tier II quality and fidelity of intervention implementation are more commonly reported when the focus is on Tier II instruction. The logic is the same as that which necessitates validation of Tier I instruction – if instruction is poor students cannot be held responsible for poor gains. In sum, it is essential that the instructional quality of the Tier I and Tier II environments be established before classifying children as poor responders and moving them into more intensive tiers of instruction.

Assessments and Criteria

As students receive intervention services, their responsiveness-to-intervention must be monitored. Researchers and practitioners currently choose among various measures of reading performance to assess RtI, including Curriculum Based Measures (CBMs; e.g. Dynamic Indicators of Basic Early Literacy Skills or DIBELS, AIMSweb; endorsed by Fuchs, 2003 as measurement of RtI), standardized measures (e.g. WRMT-R; Woodcock, 1989), and other published reading assessments. In a examination of RtI measures used in 2003-2004 for screening and progress monitoring purposes in 41 schools across 16 states, Mellard, McKnight, and Woods (2009) reported that schools used numerous and distinct instruments to measure progress in Tier II interventions,

including: published reading assessments that were not part of an intervention or core curriculum program (33%), curriculum based measures (19%), DIBELS (13%), and published reading assessments in programs designed to supplement instruction for struggling readers (12%). Recently, researchers reported using similar measures to assess students' RtI (e.g. Vaughn et al. 2007; O'Connor et al., 2005).

In addition to considering which assessments should be used to measure RtI, criteria must also be established for each assessment to distinguish among good and poor responders. Students who fall short of the criterion on an assessment or set of assessments are recognized as poor responders and in need of more intensive intervention or special education services. Several responsiveness criteria have been explored among RtI researchers, including: Low Achievement (LA), Final Benchmark, Normalization, Low Growth, Median Split, and Dual Discrepancy (DD). Each method for determining poor response is described below:

Low Achievement (LA). A student is identified as a poor responder under the LA criterion if his score on some outcome measure falls below a pre-specified cut-point. Cut-points are often based on percentiles scores (e.g., students scoring below the 30th percentile on the outcome are designated as poor responders; Al Otaiba & Fuchs, 2006). Final Benchmark and Final Normalization are variations on the LA criterion.

Final Benchmark. The final benchmark criterion requires that a student's score on the outcome after receiving Tier II intervention falls below the cut-point as specified by norm-referenced criteria (Good, Simmons, & Kame'enui, 2001). For example, Al Otaiba and Fuchs (2006) used final benchmark of 40 words correct per minute (wcpm) on

ORF to assess Tier II RtI for 1st grade readers. Vaughn, Wanzek, Murray, Scammacca, Linan-Thompson, and Woodruff (2009) also used final benchmark on ORF (i.e. fewer than 27 wcpm) to identify second grade students who responded poorly to a Tier II intervention.

Normalization. The Normalization method identifies students as poor responders when, after receiving Tier II intervention, they do not meet or exceed a set standard score (typically 90; about the 25th percentile) on some standardized test of reading (e.g. Torgesen et al., 2001).

Low Growth (LG). The LG criterion classifies poor responders as those students who have pre-to-post intervention growth that falls below that of their peers by some magnitude, which may vary. For example, Al Otaiba and Fuchs (2006) used a cut-off that identified the bottom 30% of intervention students on the amount of pre-to-post treatment growth on letter-sound and segmentation fluency measures to classify poor responders. The LG criterion has also been used with a set growth expectation (e.g. 1.5 words per week growth on ORF for 2nd graders; Fuchs et al., 2004); students who do not make growth commensurate with the expectation are identified as poor responders.

Median Split. The Median Split method is similar to the LG method. Median split is accomplished by rank ordering slopes for all students receiving Tier II intervention, finding the midpoint, and then designating all students whose slopes fall below the midpoint as poor-responders. Vellutino et al. (1996) used this method to designate students as good and poor responders to an intensive Tier II reading

intervention. By design, this method will classify approximately half of all students receiving Tier II intervention as poor responders.

Dual –Discrepancy. The Dual Discrepancy (DD) approach combines LA and LG criteria. Students qualify as poor responders if their final score is below some a-priori specified cut-point *and* if their slope is inadequate compared to other Tier I or II recipients. Fuchs and Fuchs (1998) explored cut-points for identifying dually-discrepant readers and found that designating readers whose final scores and growth were 1.0 standard deviation below that of their classmates yielded reasonable prevalence and false-negative rates. This method eliminates identification of those students who have adequate growth in response to intervention, but due to extremely low initial scores end up with a final score below the cut-point (i.e. "low achievement" qualifiers). It also eliminates identification of those students with initial scores that are near the high end of the qualification criteria and whose growth falls below average growth for Tier II students while their scores exceed the cut-point after intervention. Therefore, the DD approach is hypothesized to reduce the number of false-positives by requiring that poor-responders show low achievement and low growth, despite receipt of high-quality Tier II intervention.

Altogether, to assess a student's RtI interventionists must evaluate the quality of Tier I and Tier II instruction, choose measure(s) to monitor progress, and establish criteria for selecting students who may need more intensive intervention (i.e. the poor-responders). Just as measures and criteria used for monitoring RtI vary across studies, the pairing between measure and criteria also varies. For example, the DD criterion may be

matched with a word-reading measure to indicate RtI in one study, and the same DD criterion may be matched with an ORF measure to indicate RtI in another study. Several researchers have compared the validity of measure/criteria pairings to identify those with the greatest promise for appropriately identifying children with RD within an RtI framework. For example, Speece and Case (2001) explored the DD criterion for designating 1st and 2nd grade students as good or poor responders to Tier II intervention. The researchers designated students as poor responders if their post-test reading scores (level) and growth (slope) across the year (with at least 10 probes) on ORF probes was 1 standard deviation or more below their classmates. They varied this approach with a LA approach and an IQ-D approach for identifying RD. Results indicated that the DD approach identified children with poorer phonological processing skills and those whose academic and behavioral competencies were rated lower by their teachers, supporting the construct validity of the ORF/DD method of identifying children with RD. Furthermore, the DD group reflected the demographic and gender distributions in the population. As a result, Speece and Case concluded that compared to the LA and IQ-D approaches, the DD approach seemed more appropriate for the identification of children with RD.

More recently, researchers have been using a variety of measure/criteria combinations to judge responsiveness and have found measure/criteria combinations to disagree on responder status and to have questionably stability. In a three-year longitudinal study of RtI, Simmons, Coyne, Kwok, McDonagh, Harn and Kame'enui (2008) measured students' response to a Tier II reading intervention using a combination of CBM and standardized reading assessments. In first grade, students' RtI was

measured using the Phoneme Segmentation Fluency and Non-word Fluency subtests of the DIBELS. Students were administered the WRMT-R Word Attack, Word Identification, and Reading Comprehension subtests to measure RtI in the fall and spring of grades 1-3. Lastly, a test of ORF was given in the fall of 1st grade and fall and spring of 2nd and 3rd grades. The 30th percentile was used as the criterion on all measures, above which students were considered "out of risk" post intervention. Using criteria paired with WRMT, Simmons et al. reported consistency in risk status once children were identified as "out of risk"; that is, after receiving intervention most children whose post-intervention standard scores on the WRMT subtests exceeded the 30th percentile continued to stay out of risk on the WRMT through third grade. This was true for 93% of students when passage comprehension was used to indicate risk, approximately 95% of students when word attack scores indicated risk, and approximately 88% of students when word identification scores indicated risk. This finding seemed to demonstrate relative consistency in risk status as demonstrated by the WRMT/30th percentile measure/criteria combination for children who received early intervention, though Simmons and colleagues agreed with Fuchs et al. (2004) in that the WRMT may inflate reading scores for young children. When employing ORF scores to indicate RtI, far fewer children attained out-of-risk status and many of those who did fell back into risk later on. Simmons et al. suggested that fluency skills might require more intensive remediation than other early reading skills, which may have contributed to the disagreement between the WRMT and ORF measures on responder status (in addition to

the inflation of scores on the WRMT). In any case, group designation and stability of designations differed depending on the referenced measure.

Some researchers have more directly investigated the issue of disagreement in RtI status when using various measurement and criteria to judge RtI. For example, Barth and colleagues (2008) investigated agreement among RtI criteria for determining responsiveness to Tier II intervention and found poor agreement overall. The researchers compared slope, intercept and DD criteria for the Sight Word Efficiency and Phonemic Decoding Efficiency subtests of the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1997) as well as ORF measures. The progress monitoring assessments were also administered at end-of-year along with the Word Attack, Letter-Word Identification, and Passage Comprehension subtests of the Woodcock Johnson III to measure final reading proficiency. All possible measures and criteria were explored over a range of cut-points to classify groups of children as good or poor responders. Results indicated that good and poor responder grouping agreements overwhelmingly depended on the cut-points applied to RtI criteria. Also for end-of-year measures, those that assessed similar constructs had the highest degree of agreement regarding students' final reading status; however, agreement also increased when similar cut-points were used which may have accounted for some of the agreement among measures (Barth et al., 2008). The researchers concluded that the cut-points and measures used to classify good and poor responders drove agreement among group designations, with no clear preference for intercept, slope, or DD criteria for identifying good and poor responders. Nevertheless, the authors supported the LD summit consensus

(Bradley, Danielson, & Hallahan, 2002) to use growth measures to determine RtI and to rely on highly reliable norm-referenced assessments to indicate reading proficiency.

Although exploring the group membership agreement across measure/criteria combinations and over time may help to identify measures and criteria that are more reliable and stable, identifying measures and criteria for RtI that are predictive of future reading proficiency is essential if RtI is to be used to classify children as RD. To this end, Burns and Senesac (2005) investigated the capability of the DD criterion to differentiate end-of-year reading proficiency for 1^{st}-3^{rd} grade readers who had received Tier II intervention. Students whose teachers identified them as "poor readers" and who scored below the 25^{th} percentile on their state reading exam were selected for intervention. Participants received one of two Tier II interventions after which their reading proficiency was established using Gray's Oral Reading Test (GORT, Wiederholt & Bryant, 2001). DIBELS ORF was used as the RtI measure; post-intervention risk status was determined using grade-specific published benchmarks and intervention responsiveness was determined by applying 4 levels of severity (25^{th} percentile, 33^{rd} percentile, 50^{th} percentile, and <1 standard deviation below the mean) as cut-points to mid- to end-of-year ORF growth scores. Therefore, students whose final ORF score fell below the published grade-specific DIBELS cut-point for at-risk status and whose growth in reading fell below the relevant severity cut-point were considered dually discrepant. DD and non-DD groups were compared on their end-of-year GORT scores to determine whether post-intervention reading scores were significantly different between groups. Results indicated that end-of-year GORT reading scores for DD and non-DD groups were

significantly different when each percentile cut-point (but not on the 1 standard deviation cut-point) was used to determine the growth component for DD status. Effect sizes, which reflected differences in GORT scores between DD and non-DD groups, were greater than 0.79 for all percentiles and was 0.64 for the 1SD criterion. This first step validated the use of DD to predict RD when paired with 25th, 33rd, or 50th percentile cut-scores to indicate adequate growth.

Post-intervention GORT scores were also compared between students who were considered at-risk, some-risk, and out-of-risk according to DIBELS benchmark criteria. Results indicated that end-of-year reading proficiency based on GORT mirrored end-of-year risk-status based on ORF benchmarks. That is, at-risk students scored significantly lower on GORT than some-risk students, who scored significantly lower than out-of-risk students. Next, GORT scores were compared for students who did and did not make adequate growth according to the three percentile criteria and the 1SD criterion. GORT scores were statistically distinct under the percentile cut-point criteria, with effect sizes for good and poor responder groups ranging from .54 (50th percentile cut-point) to .69 (25th percentile cut-point). Use of the 1 SD criterion to indicate adequate growth did not result in distinguishable GORT scores (effect size = 0.51). Finally, GORT scores were compared within the subset of students who were at-risk post-intervention, but whose *response* to (growth during) intervention differed. This group included students who had very low initial scores and made adequate growth (poor readers, good responders) and students with higher initial scores who made very poor growth (poor readers, poor responders). GORT scores were not statistically distinguishable for students within the at-

risk group at any growth cut-point and effect sizes were small to moderate. This indicates that within the group of poorest readers, growth estimates may not effectively identify good and poor responders groups who are distinguishable on post-intervention outcomes.

Prevalence rates for DD readers within the tutored sample and the student population (all 1st-3rd graders in participating schools) were also estimated. Prevalence rates for students identified as DD after receiving intervention out of all tutored students were: 10.8% for 25th percentile and 14.9% for the 33rd percentile and 29.7% for the 50th percentile and 12.3% for the 1 SD criteria. Prevalence rates for DD readers within the student population were: 3.7% for the 25th percentile, 4.4% for the 33rd percentile, 6.6% for the 50th percentile, and 1.9% for the 1 SD below the mean criteria. Lerner (2003) estimated the national prevalence of LD to be about 5% of the school-based population. As Burns and Senesac point out, the 33rd percentile criterion for RtI most closely matched national prevalence estimates. However as in other studies, group designation of students with good and poor response and the ability of measurement/criteria combinations (as measured by effect size) to predict future reading proficiency changed as a function of RtI criteria.

A limitation of Burns and Senesac (2005) is that the researchers focused on only ORF as a measure of RtI and GORT as an outcome. Although each of these measures are commonly used in public schools for diverse purposes, the validity of each for measuring reading skills of young readers has been questioned, especially when the tests are used in isolation. Given the diversity of RtI and standardized assessments available to

practitioners, it is important to explore the relationships between several RtI and outcome measures within and across samples. Fuchs et al. (2004) partially accomplished this task. In their study, Fuchs et al. explored whether Tier II Progress Monitoring (PM) measures adequately predicted end-of-year reading proficiency for 1st and 2nd grade readers. Students with poor response to Tier I reading intervention (PALS) received supplemental Tier II reading intervention in small groups. For the 1st graders, the Median Split criterion was applied to weekly Word Identification Fluency (WIF) and DIBELS Nonsense Word fluency probes; these measures were used for PM. The Normalization criterion was applied to post-intervention WRMT scores (i.e. responsiveness was determined relative to a standard score of 90) and the Final Benchmark criterion was applied to the DIBELS ORF measure (i.e. students scoring below the benchmark for end of 1st grade – 40wpm – were designated as poor responders) as additional indicators of students' RtI. For each RtI measure and criteria, Fuchs et al. explored the validity of designated good and poor responder groups by determining the extent to which measure/criteria combinations differentiated between students' end-of-year performance on 5 outcome measures: 1) standard scores on the WRMT-R Word Attack; 2) standard scores on the WRMT-R Word Identification; 3) spelling standard scores on the WIAT; 4) DIBELS ORF; and 5) comprehension raw scores on the Comprehension Reading Assessment Battery. Thus, good and poor responder groups were established using each of the 4 RtI measures/criteria in turn and end-of-year reading scores on the 5 outcome measures were compared between groups to determine which measure/criteria best differentiated between students based on end-of-year outcomes.

Overall, RtI measure/criteria combinations did not agree on good and poor responder group status and differed in the ability to discriminate between proficient and non-proficient readers on outcome measures. Virtually all intervened students (8.4% of the full sample) were classified as poor responders under the Final Benchmark (ORF) criterion as compared to the 1.4% so classified by the Normalization method. For each of the two slope measures, the Median Split method identified approximately half of the intervened sample as poor responders (3.5% of the full experimental sample). Measures of effect size for the ability of RtI measure/criteria combinations to differentiate between proficient and non-proficient readers were also provided. The Normalization method discriminated between good and poor readers on 4 of 5 reading measures and 2 of 3 progress monitoring measures, with average effect sizes of 1.59 for outcome and 1.05 for growth. Effect sizes were large for end-of-year level (ES = 1.0) and growth (all ESs >.90) on WIF and moderate for the Nonsense Word fluency measure (average ES=0.46). When comparing the Normalization method to the WIF slope method, WIF slope method fared better, with significant effects on every measure.

For 2nd graders, RtI was judged in 6 ways: Median Split criteria applied to 1) Woodcock word-reading gain scores and 2) CBM slope; 3) post-treatment Normalization (standard score of 90 or above) criteria applied to WRMT word-reading; 4) post-treatment Final Benchmark criterion applied to ORF; 5) a normative criterion for expected ORF slope at grade 2 (1.5 words per week increase); and 6) Dual Discrepancy criterion applied to the ORF slope at grade 2 and Final Benchmark for ORF for grade 2 (i.e. at least 75 wcpm at end of treatment). Reading measures used to derive end-of-2nd

grade outcomes were the same as those used for 1st graders. The CBM slope (method 2) and Dual Discrepancy (method 6) identified identical groups of children as good and poor responders. These methods differentiated between proficient and non-proficient readers with better consistency and larger magnitude of differences compared to other methods, with average effect sizes of .85 for outcome variables and .84 for growth (Fuchs et al., 2004). Effect sizes on measures of reading comprehension were above 1.0 for each measure/criteria combination. Overall, Fuchs et al. demonstrated that the measure/criteria combinations used to designate RtI status often resulted in distinct groups of students classified as poor responders. Furthermore, some measure/criteria combinations for assessing RtI were better predictors of later reading achievement as demonstrated by effect size differences.

Fuchs et al.'s (2004) study was limited in terms of grade-levels explored and longevity of prediction -- students' RtI scores were used to predict reading proficiency at the end of their current school year and not beyond. O'Connor et al. (2005) and others (Keiffer, 2010; Simmons et al., 2008) report fluctuations in reading performance for students who receive Tier II interventions as they move through school; students who exit intervention in one grade may be found eligible for additional intervention in later grades when reading becomes more difficult. Therefore, it is important to follow students over more than a single year to establish the longevity of their RtI and whether early reading measures can predict RD in later grades when the nature of RD may change.

In 2008, Fuchs et al. conducted a study that followed students from grade 1 to grade 2 and explored the validity of RtI measure/criteria combinations for predicting

long-term RtI. The research was initiated by the Office of Special Education Programs (OSEP) and carried out through the National Research Center on Learning Disabilities (NRCLD). One of the three study focus questions was how to define responsiveness and non-responsiveness as a means to provide the basis for distinguishing RD from non-RD. As found previously, Fuchs et al. (2008) determined that measure/criteria combinations based on 1st grade reading scores obtained from students participating in the second tier of an RtI program differed in the selection of students as good and poor responders. To make sense of these disagreements, the researchers extended their investigation to determine which identification methods had superior sensitivity and specificity rates when predicting students who truly developed RD by the end of second grade. A student was designated as RD at the end of 2nd grade if s/he scored more than 1 standard deviation below the national mean on a composite measure of Test of Word Reading Efficiency (Sight Word Efficiency) and on Woodcock Reading Mastery Tests (Word Identification, Word Attack, and Passage Comprehension). First grade curriculum-based measurement probes for Word Identification Fluency (WIF; Fuchs, Fuchs, & Compton, 2004) were coupled with 6 RD identification methods including: final IQ-D, initial LA, final Normalization, Final Benchmark, Slope Discrepancy, and DD to determine which method proved as the best indicator for RD (Fuchs et al., 2008).

Results indicated that the IQ-D criterion was inadequate as a tool for identifying RD when paired with the WRMT; despite good hit rates and specificity, this method provided "unrealistically low prevalence" and poor sensitivity (Fuchs et al., 2008, p. 432) when used to identify 1st graders who would fail to meet performance expectations by the

end of 2nd grade. Given the tendency for the WRMT to inflate the reading scores of early readers (Fuchs et al., 2004; Simmons et al., 2008), this finding is not too surprising. To assess the validity of RtI measures for identifying RD, Fuchs et al. applied decision rules (including specificity of at least 0.80 and sensitivity of 0.80) to the RtI assessments to identify the most suitable response criteria. Several 1st grade measure/criteria combinations produced adequate probabilities for predicting RD, including: (1) initial LA (< 1 SD mean) on WIF; (2) final Normalization (standard score < 90) using TOWRE Sight Word Efficiency; (3) Slope Discrepancy using WIF (at least 1 SD below a normative sample); and (4) DD using passage fluency for level (with a score of less than 40wpm as the cut-point) and WIF for slope (at least 1 SD below normative sample).

Overall, research suggests that different RtI measures/criteria have poor agreement when classifying students as good or poor responders to instruction in Tiers I and II (e.g. Barth et al., 2008; Brown-Waesche, 2011; Fuchs et al., 2004, 2007; Speece & Case, 2001). Additionally, when classifying good and poor responders in response to Tier I instruction, classifications tend to be unstable over time (Brown-Waesche et al., 2011; Francis et al., 2005). Research seems to suggest that the best approaches for judging RtI include use of growth and final achievement measures (i.e. the Dual Discrepancy criterion) paired with validated CBMs and/or standardized assessments. Even still, there are a range of RtI assessments available and researchers and practitioners differ in their preferred measure (Mellard et al., 2009). So due to the wide range of RtI measures available, the different criteria and cut-points applied to measures to determine responsiveness to intervention and the tendency of measure/criteria combinations to

disagree regarding a student's RtI, it is important to establish whether there are some measure/criteria combinations that can more reliably distinguish between those children who will continue to struggle despite intervention and those for whom intervention will sufficiently remediate their deficits. Fuchs et al. (2004) made progress toward this end when they identified measures and criteria that differentiated between good and poor responders by end-of-year for 1st and 2nd graders, with large effect sizes. However, Fuchs et al. only predicted reading scores through the end of the same year in which intervention was received. In another study, Fuchs et al. (2008) predicted reading proficiency for 1st graders through the end of 2nd grade. First and 2nd grade students encounter few words that are non-decodable and instruction is still mainly focused at the word level (O'Connor, 2007). In fact, O'Connor, Bell, Harty, Larkin, Sackor, and Zigmond (2002) estimated that approximately 60%-70% of words read in a text at the 1-2nd grade level are seen in other texts at the same level, but percent of overlap drops to about 50% in 3rd – 4th grade (see also O'Connor, Swanson, & Geraghty, 2010). Therefore, higher percentages of unique words per page and longer, more complex sentence structures found in text beyond 2nd grade make reading more difficult. Also, older students are required to decode multisyllabic words to be successful readers (Lenz & Hughes, 1990; O'Connor, 2007) and there is a greater focus on reading comprehension (Klingner, Allison, & Boardman, 2007) beyond 2nd grade. Due to these changes, students who are proficient readers in 2nd grade may struggle to reach proficiency in later grades, thereby possibly exhibiting late-emerging RD (Catts, Compton, Tomblin, & Bridges, 2011). In addition, students tend to be referred for special education services most often

between their 3rd and 4th grade years. Therefore, it is essential to establish whether responsiveness status determined by 1st and 2nd grade measures remains an accurate predictor of RD in 3rd grade and beyond given the possibility of additional effective instruction.

Undoubtedly, it is important to extend research on the use of the RtI framework to predict RD. Researchers should investigate whether the same (or different) measures and criteria found predictive of RD through 2nd grade in earlier studies can accurately distinguish between students who will be reading proficiently in later years and those who will continue to struggle through 3rd grade and beyond. This dissertation builds upon Fuchs et al. (2008) by extending the prediction of RD by an additional year – that is by 3rd grade. Also, since previous researchers identified their tutored samples in 1st grade without allowing for students to enter intervention if they became eligible later (Fuchs et al., 2008; Simmons et al., 2008), it is impossible to determine whether their findings generalize to students whose reading difficulties emerge later in schooling. Therefore, this research requires replication with a new sample that includes students who become eligible for intervention in later grades.

Using a longitudinal approach, I identified children who, despite receiving intensive, high-quality intervention as needed from 1st through 3rd grade continued to have difficulties with decoding and/or comprehension tasks. I entered 1st and 2nd grade predictors into logistic regression analyses to determine the ability of early measures to predict later occurrence of RD (i.e. in 3rd grade). Next, I paired predictive measures with responsiveness criteria to form groups of good and poor responders, and I calculated

sensitivity and specificity indices for each measure/criteria combination for the prediction of RD. Finally, I applied the most efficient logistic regression equations used to identify group membership for an initial cohort of students on a second cohort. I repeated sensitivity and specificity analyses of measure criteria combinations on the second cohort to determine the extent to which results replicated.

Chapter 3: Method

Setting

Data were collected from two Southern California school districts, henceforth referred to as District A and District B. A summary of demographic information for each District is provided in Table 1. In the 2007-2008 school year, District A served over 56,000 students. Most students (approximately 85%) identified as ethnic minorities. The largest ethnic subgroup was Hispanic (68.1% of the student body), followed by African Americans (16.3%) and Whites (10.9%). Students who identified as American Indian/Native Alaskan, Asian, Filipino, and Pacific Islander each comprised less than 1% of the student population. Approximately 73% of students were socioeconomically disadvantaged, 43.8% of students were English Language Learners (ELLs), and 9.2% of students were enrolled in Special Education programs. District B served approximately 20,000 students, most of whom were ethnic minorities. Like District A, the largest ethnic subgroup were Hispanic (71.8%), followed by Whites (15.4%) and African Americans (4.4%). American Indian/Native Alaskan and Pacific Islanders comprised the smallest proportion of the student body, each at less than half a percent. Approximately 65% of

the student body was socioeconomically disadvantaged, half were ELLs, and 8.7% were enrolled in Special Education.

Students attended one of five elementary schools across Districts A and B (Table 2). School A, School B, and School C belong to District A and School Y and School Z belong to District B. In 2007-2008, schools were comparable in size, except for school A which served approximately twice as many students as the remaining schools. On average, ethnic representations of the student bodies for each school reflected those estimates for the district. In all schools, Hispanic students comprised the largest proportion of the student body. ELLs comprised between 30 and 60 percent of students at each school; approximately 95% of ELLs at each school spoke Spanish as their first language. Special education information was not available by individual schools, but district information was available (see above).

Classroom Instruction

For the larger RtI study, teachers at each of the 5 schools participated in 120 hours of language arts professional development in reading as part as two California mandates: Assembly Bill 1485 (AB1465) The Reading First California Technical Assistance Center and Assembly Bill 466 (AB466) for under-performing districts. Professional development was delivered by approved providers, and was divided between 40 hours of direct class instruction to teachers and 80 hours of follow-up. Participating teachers attended Reading First Institutes, all of which stressed development of Reading First components of phonemic awareness, phonics, vocabulary, fluency, and comprehension. Professional development materials and procedures were authored and published by the

California Reading Development Center, the Sacramento Reading Implementation Center (RIC) and the Reading Lions Center at the Sacramento County Office of Education. Professional development focused on use of the state adopted Houghton Mifflin Language Arts series and accompanying resources, which were accepted as evidence-based reading instruction in California's application for Reading First funding. Instructional strategies included direct instruction, guided reading, frequent diagnostic assessments, providing for universal access (small group instruction and differentiated instruction), and a pacing guide for coverage of the California Language Arts standards. Principals at these schools were required to incorporate these elements in teacher evaluations and classroom observations.

Participants

Participants included 418 1st graders who were present at one of the five study schools from 1st to 3rd grade. Participants were initially recruited as part of a Response to Intervention study that involved students from kindergarten to 4th grade. Cases were included in the current sample if the following conditions were met:

1) Participants were 1st graders in 2007-2008 or 2008-2009,
2) Participants had *access* to Tier II intervention from 1st to 3rd grade,
3) Participants had complete or near complete data on all relevant predictors and outcomes.

Condition 1 was necessary given the intent of the current study. For Condition 2, access is defined as having the opportunity to participate in Tier II intervention given qualifying test scores during universal screenings. Including data for participants who were denied

access to intervention for any reason (e.g. parent declined consent, student was wait-listed, etc.) would undermine efforts to identify RtI measure/criteria combinations that most accurately predict 3rd grade reading proficiency or RD. Given attrition common to longitudinal studies, Condition 3 ensured that all included cases had sufficient data to support proposed analyses.

Of the 418 participants who met inclusion criteria, 31 cases were removed due to missing data on all outcome and/or more than half of the predictor variables (See Data Analysis). The remaining 387 cases were split into two cohorts: students in Cohort A (*N*=219) attended 1st – 3rd grades from 2007-2010 and in Cohort B (*N*=168) attended 1st – 3rd grades from 2008-2011. The purpose of dividing the sample was to enable exploration of the extent to which results generated from the initial prediction models created using data from one cohort could be replicated with data from the other. Cohort specific participant descriptions are provided below.

Cohort A. Sample identification and missing data handling procedures resulted in the inclusion of 219 cases subject for analyses from Cohort A. All participants attended 1 of the 5 target schools from 1st – 3rd grade and had access to Tier II intervention as needed each year. A summary of descriptive statistics for Cohort A participants is provided in Table 3. Approximately half (*N*=108) of the participants in Cohort A are male. The participants largely identify as ethnic minorities, with Hispanics comprising 71% of the sample, followed by African Americans (12%) and Whites (11%). About half of participants were ELLs in 1st grade with an average score on the California English Language Development Test (CELDT) score of 3, which indicates an

"intermediate" level of English proficiency (on a scale from 1 "beginning" to 5 "advanced"). Four students were identified for special education during the study and moved into Tier III of a three Tier response-to-intervention framework. Their scores were retained in analyses.

Cohort B. Sample identification and missing data handling procedures resulted in the inclusion of 168 cases subject for analyses from Cohort B. All participants attended 1 of the 5 target schools from 1st – 3rd grade and had access to Tier II intervention as needed each year. A summary of descriptive statistics for Cohort B participants is provided in Table 3. Similar to Cohort A, approximately half of participants in Cohort B are male. The majority of students identify as Hispanic (77%), White (10%) or African American (7%). Approximately half the participants were ELLs in 1st grade with an average CELDT score of 3.02, which reflects an "intermediate" level of English proficiency. Four students were identified for Special Education during the study and moved into Tier III; their relevant reading scores were retained for analysis.

First grade PPVT scores were unavailable for Cohort B. In 3rd grade, the sample average PPVT standard score for Cohort B (M= 87.95, SD=10.78) was statistically equivalent to that of Cohort A (M=86.2, SD=11.44, p=.12; Table 3). The sample average TOLD-P:3 Relational Vocabulary standard score in 1st grade was approximately 1 point higher for Cohort B (M=8.3, SD=3.0) compared to Cohort A (M=7.3, SD=3.6, p=.008). According to these scores, participants from each cohort scored below the national normed average (M=100, SD =15) on PPVT in 3rd grade, and on the TOLD (normed M=10, SD=3) in 1st grade.

Measures

Assessments for sample selection and description. The Peabody Picture Vocabulary Test-3rd edition (PPVT-R; Dunn & Dunn, 1997) was used to describe receptive language in English for all students. The PPVT is an individually administered, norm-referenced measure of receptive vocabulary designed for individuals 2.5 years old through adult. The child selects from among four pictures, one which best represents a word read by the examiner. Standard quotient scores are reported here (raw scores standardized for age in years and months at the time of testing), with a mean of 100 and standard deviation of 15.

Subtests of the Dynamic Indicators of Basic Early Literacy (DIBELS; Good & Kaminski, 2003) and Word Identification Fluency (WIF; Fuchs et al., 2004) were used for selection for intervention and monitoring the students' responsiveness (see descriptions below). All tests were timed and individually administered.

First grade measures. First grade students received six individually administered assessments: Letter Naming Fluency (LNF), Phonemic Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), and Oral Reading Fluency (ORF) from the DIBELS battery of assessments, Word Identification Fluency (WIF; Fuchs et al., 2004) and the Test of Language Development 3rd edition (TOLD; Newcomer & Hammill, 1997). The WIF, ORF, and relational vocabulary subtest for the TOLD were included in analyses for this study and are described below. WIF was administered in the fall, winter, and spring; ORF was administered in the winter and spring; and TOLD was administered in the winter.

The WIF assessment consists of word lists developed by Fuchs et al. (2004), which contain 100 isolated words randomly selected from Dolch pre-primer, primer, and first grade high frequency word lists. Students are presented with this list and asked to read the words as quickly as they can. The score is the number of words read correctly within one minute, a measure of automaticity of reading skill. Fuchs et al. (2004) report alternate test form reliability of .91 from two consecutive months. The alternate-test form stability coefficient for two consecutive weeks was .92 in the beginning of first grade.

DIBELS ORF measures reading rate and accuracy. ORF passages were administered in the winter and spring for first graders and in the fall, winter, and spring for second graders. ORF was also administered every three weeks to students in the Tier II interventions to monitor progress and as an outcome in spring of each year. The DIBELS ORF passages are used nationally extensively to identify students who need instructional support and to monitor educational progress. Benchmark passages are available for each grade level for fall, winter and spring of each year. The student is presented with three different passages and asked to read aloud for a period of one minute for each. Scores are calculated as the number of words attempted minus errors, and the median score is used for analysis. Alternate-form reliability ranges from .79 to .94 across measures, and inter-rater reliability for the first grade sample is 0.95.

TOLD- I:3 is an individually administered, norm-referenced measure with established reliability and validity (Newcomer & Hammill, 1997). The relational vocabulary subtest measures a child's ability to understand and orally express the

relationship between a pair of spoken words, without picture cue. A 1 is recorded for each correct response and 0 for each incorrect response; the maximum raw score for relational vocabulary is 30. Standard scores are reported here, with a mean of 10 and standard deviation of 3 for the relational vocabulary subtest.

Second grade measures. The DIBELS ORF subtest (described earlier), TOLD: I-4 (Hammill & Newcomer, 2008), and the Word Identification (WID), Word Attack (WATT), Word Comprehension (WC), and Passage Comprehension (PC) subtests of the Woodcock Reading Mastery Tests - Normative Update (WRMT-NU; Woodcock, 1998) were administered to all second grade students. ORF was administered in the fall, winter, and spring of 2nd grade, TOLD was administered in the winter, and the WRMT subtests were administered in the fall. These measures served as predictors of reading disability in logistic regression analyses.

TOLD I:4 is an individually administered, norm-referenced measure with established reliability and validity (Hammill & Newcomer, 2008). The relational vocabulary subtest measures a child's ability to understand and orally express the relationship between three spoken words, without picture cue. A 1 is recorded for each correct response and 0 for each incorrect response; the maximum raw score for relational vocabulary is 30. Standard scores for the relational vocabulary subtest are reported here, with a mean of 10 and standard deviation of 3.

The WRMT-NU WID subtest requires students to identify words in isolation; the WATT subtest requires students to apply phonic and structural analysis to pronounce pseudowords; the WC subtests require students to identify analogies to written words;

and the PC subtest requires students to read 1 or 2 sentences silently with a missing word signaled by a blank space, and to supply a word that made sense in that space. Standard scores for each subtest were used in analyses unless otherwise noted (M=100; SD=15).

Outcome measures. ORF, PPVT, and subtests of the WRMT-NU (Woodcock, 1998) were used as outcome measures at the end of 3rd grade (see previous descriptions). The Test of Written Spelling (TWS-4; Larson, Hammill, & Moats, 1999) was also used as an outcome.

The TWS-4 (Larson et al., 1999) is a norm-referenced, group-administered spelling assessment. There are two alternate equivalent forms, each of which contains words with predictable and unpredictable spellings. Administration proceeds by dictating 20 progressively difficult words to students. The TWS-4 publishers report high coefficient alpha, test-retest and inter-rater reliability, with all coefficients consistently greater than the standard of .9 (Salvia & Yssledyke, 1998). The final score on the TWS-4 is the number of words spelled correctly.

The English Language Arts (ELA) section of the California Standards Test (CST) was used to further describe proficient and RD reader groups (see Procedures). Students in grades 2-11 take the CST as part of the Standardized Testing and Resorting (STAR) Program; the test was designed to measure students' progress toward mastering the California state academic standards. Descriptions of the CST and sample test questions were derived from the California Department of Education website (www.cde.ca.gov). The ELA portion of the CST for third grade students contains approximately 65 questions that assess students' skill in word analysis, reading comprehension, literary response and

analysis, writing strategies, and writing conventions. The word analysis strand assesses students' ability to decode words and derive word meaning from word and text structure and content. A sample word analysis question is "How should the word 'chambers' be divided into syllables?" followed by four answer choices "(A) cham-b-ers, (B) cham-bers, (C) ch-am-bers, (D) cha-mbers." The reading comprehension portion of the ELA subtest requires students to read a page-long selection of text and use comprehension strategies (prediction, comparing information from two or more sources) to answer passage related questions. Questions require straight text recall, use of prior knowledge, understanding of main idea and supporting details, understanding of problem-solution text structures, and ability to follow multi-step directions. Comprehension questions are presented alongside the relevant story. A sample text-related comprehension question is: "What did Monkey do as soon as the dogs became bored and went away?"

Scores for each section of the CST (i.e. English Language Arts, Math, Science) range from 150 to 600. Five levels of proficiency are associated with specific score ranges: Far Below Basic (FBB) scores range from 150 to 258; Below Basic (BB) scores range from 259-299; Basic scores range from 300-349; Proficient scores range from 350-401; and Advanced scores range from 402-600. Students scoring FBB or BB on a particular subtest demonstrate extremely limited or flawed knowledge of the content evidenced by overall poor performance; students scoring Basic demonstrate a partial or rudimentary understanding of knowledge and skills assessed; students scoring Proficient demonstrate a competent and adequate understanding of knowledge and skills assessed; and students scoring Advanced demonstrate a comprehensive and complex understanding

of knowledge and skills assessed (Table 4; California Department of Education, http://www.cde.ca.gov).

Procedures

Intervention eligibility. Participating students completed universal screening in their general education classes in the fall, winter, and spring of each year. Students who met intervention eligibility criteria (see below) were referred to small group intervention. Students could be referred for intervention at any time point (fall, winter, or spring) given they met intervention criteria and had attended one of the five schools in 2007-2008. Eligible students received intervention in a Tier II setting with trained researchers and graduate students until they met pre-specified exit criteria across two consecutive time-points. Special education services supplanted Tier II intervention if students were referred during the course of the study; however, progress monitoring data continued to be collected so these students' scores (n = 4 for Cohort A; n = 4 for Cohort B) are included in analyses.

Students were selected for Tier II intervention if their fall of 1st grade scores fell below specified cut-points on WIF (<8) or on any of the LNF (<45), PSF (<30), NWF (<25), or Oral Reading Rate (<1) DIBELS subtests (Kaminski & Good, 1996). Cut-points differed from those recommended in the DIBELS manual due to earlier studies (O'Connor, 2005; O'Connor et al., 2010) that found published criteria too low to identify nearly all (90%) students who later demonstrated poor reading achievement. ORF was used for intervention eligibility determinations for 2nd and 3rd grade students. Second grade students were selected for Tier II intervention if their median ORF score fell below

26, 52, and 70 wcpm any time during the fall, winter, and spring time points, respectively.

DIBELS measures were administered every 3 weeks for progress monitoring and scores were used to identify students whose performance warranted exit from intervention. Exit criteria for 1st grade required students to have scores greater than 35 on PSF in fall, winter, or spring; scores greater than 30 on NWF in the fall or greater than 50 in the winter and spring; or scores exceeding 20 on ORF in winter or 40 wcpm in spring. The 2nd grade exit criterion was based solely on ORF. To exit intervention, students' ORF scores must have exceeded 44 wcpm in fall, 68 wcpm in winter, or 90 wcpm in spring.

Tier II intervention. The Tier II intervention consisted of small group (two or three students) instruction for 25-30 minutes in first and second grade, four times per week. Intervention was offered September through April during the academic year. Students in the intervention sample were assessed every three weeks with measures of Non-word fluency (NWF), WIF, and ORF (beginning in winter of grade 1). Across grades, a gating procedure was used such that students who reached high levels of performance on lower level skills were shifted into more difficult measures.

Tier II intervention for 1st graders and for 2nd graders with low fluency scores was based on *Sound Partners* (Vadasy et al., 2005). Instruction occurred in small groups and covered letter-sound correspondence, decoding, sight word identification, and reading of sentences and decodable books. These activities have generated significant improvement for low-skilled 1st and 2nd grade students (Vadasy, Jenkins, Antil, Wayne, & O'Connor,

1997; Vadasy, Sanders, & Tudor, 2007). Second and 3rd grade Tier II instruction included word study (decoding), vocabulary, and comprehension activities, reading and rereading books at students' current reading level, and brief spelling and sentence-writing opportunities.

In each grade, students receiving Tier II who scored above the risk cut-offs on progress monitoring measures for two consecutive measurement occasions were released from intervention, but continued to be monitored throughout the year. Likewise, any student who scored below risk cut-offs on the three-times-per-year screening measures was folded into Tier II instruction. Thus the Tier II sample received intervention as needed; approximately 12% of the sample participated in Tier II continuously during the years they had access to treatment.

Intervention was delivered by project staff during the regular school day. Students were pulled from their general education classroom according to classroom schedules such that Tier II instruction supplemented, rather than supplanted, English Language Arts instruction or ELL programming.

Tutors and training. Tutors included experienced special education teachers, classroom teachers, graduate students, teacher credential candidates, and teacher assistants. Across the staff, 61% of these individuals were tutors for the entire three years of this research; 88% were with the project for two or more years. All tutors received training from the PI of the larger study in instructional delivery of the specific curricula for each grade level (*Sound Partners* for first and second grade, decodable books with fluency practice and comprehension work in second and third grade). The initial training

was four hours and included a theoretical introduction of each activity, modeling of the activity, guided practice, and independent practice in small groups with observation, feedback, and discussion of common problems. Tutors received a teacher manual generated by the PI and Co-PI, which besides the student curricula and teacher scripts, included a pacing guide for daily lessons, a pacing guide for monthly progress based on average progress, and flow charts linking specific types and levels of activities to progress monitoring benchmarks. This initial training was supplemented by bi-monthly follow up training (again taught by the PI) where new activities were introduced, common issues noted during field observations were discussed, and additional practice provided. Project staff also met bi-weekly in small groups by school site and reviewed progress monitoring data and monthly lesson plans for each individual student. These reports and plans were reviewed and approved bi-weekly by the PI, instructional staff lead, and the project director.

Tier II treatment fidelity. An experienced classroom or special education teacher was designated as the lead tutor at each site. In addition to observing, supporting and providing feedback to the project staff, the lead tutor oversaw weekly progress of students and modifications to the monthly lesson plans. The lead tutor collected daily activity logs completed by tutors for each of the small groups, and these were reviewed by PI and Co-PI. Reviewers looked for two indications of treatment fidelity when reviewing activity logs: completion of the each of the steps/activities outlined in the teacher scripts, and student growth in DIBELS measures. Flat line progress triggered a conference where activities and/or pacing were changed. Depending on the grade level,

progress through the curricula was defined as successful completion of Mastery Tests for blocks of lessons (first and second grade, *Sound Partners*), and movement through a series of leveled readers of increasing difficulty with 85% accuracy (second and third grade curricula). Daily activity logs were entered into the annual data base and included information regarding the minutes of compiled treatment, the activities completed within the lessons, and the tutor's assessment of the child's performance. These activity logs were compared with observations of tutors conducted by the PI, instructional staff lead, and the project director. The fidelity observations included specific actions on the part of the tutors for each segment of the lesson. Fidelity was computed as a percentage of all observed actions compared to all actions expected. If any observation fell below 85% fidelity, the tutor was provided coaching and feedback, followed by co-teaching with the lead tutor until acceptable fidelity was reached. The average fidelity rating for *Sound Partners* was 92.06% and for second and third grade curricula 89.4%.

Tier I treatment fidelity. Tier I reading instruction was observed three times per year in each 1st and 2nd grade classroom that contained participating students. A low-inference observation tool was developed for data recording during classroom observations. The form was designed to record minute-to-minute activity and includes three basic categories: student level of engagement, literacy skill addressed , and teacher instructional behaviors (e.g. direct instruction, modeling). The recorded data was supplemented by 5-point Likert Scale holistic rating (1 = poor; 5 = excellent) of the observed lesson, similar to the overall ratings in the ELL Classroom Observation instrument (Gersten, Baker, Haager, & Graves, 2005) and those used in previous studies

of teacher effects (e.g., Foorman, Schatschneider, Eakin, Fletcher, Moats, & Francis, 2006). Observers drew from notes taken in the classroom for making rating judgments and provided justification of the rating based on what was observed. The resulting effectiveness rating accounted for several aspects of instruction, including: instructional flow, concreteness of presentation, appropriateness and clarity of examples, use of extension and remediation, efforts to attract and maintain student attention, and in-the-moment decisions regarding direction of the lesson based on student responses. Scores on the holistic scale served as quality indicators for Tier I instruction for the purpose of this study.

Holistic teacher ratings gathered from the Likert-Scale rating tool (described above) for 1st grade teachers ranged from 1 to 4.6 with an average of 2.75. Ratings for 2nd grade teachers ranged from 1 to 5 with an average of 2.82. Ratings for 3rd grade teachers were unavailable. Although all teachers received 120 hours of professional development in best practices for teaching reading, Tier I instructional quality varied across classrooms and was "average" overall.

Definition of reading disability. The purpose of this study was to identify measures that when collected in 1st and/or 2nd grade, could predict reading proficiency and disability in 3rd grade for students with access to Tier II reading intervention as needed. Several approaches are viable for predicting reading performance in a later grade with reading scores collected during and after receiving intervention in earlier grades. One approach is to create groups of good and poor responders to Tier II intervention using scores on measures collected in early grades and to determine whether the

performance of each group on a battery of reading measures administered in later grades differ significantly (see Fuchs et al., 2004; Burns & Senesac, 2005). The focus in this approach is on the initial identification of good and poor responders to intervention and their subsequent reading performances, rather than the identification of proficient and RD readers in later grades. Although this method may reveal plausible definitions of good and poor responders, it has limited ability to control whether average performance on outcomes for the poor responder group is indicative of reading disability, even if performance is statistically lower than that of the good responder group.

An alternative approach is to form groups of proficient and RD readers based on reading performance in later grades (e.g. grade 3) and then to use scores on earlier measures to predict later group membership for students who had access to intervention from 1st to 3rd grade. Versions of this approach have been used by Catts et al. (2011) and Vellutino, Scanlon, Zhang, and Schatschneider (2008). The latter method allows control over the definition of "proficient (or not) reading" in later grades and also preserves the ability to use continuous measures in earlier grades to predict a student's group membership.

Under the alternative approach, a composite score that reflects skill in decoding, fluency, and comprehension might be used to describe students' reading skill in 3rd grade; alternatively, children might be identified as proficient or RD readers in a single area of reading. The 'reading composite' definition allows a comprehensive picture of reading development, but it also may also obscure deficits that lie primarily in one area of reading development. In a study of late-emerging poor readers, Catts et al. (2011) found that

52% of late-emerging poor readers had delays in reading comprehension alone, 36% had delays in word reading alone, and 12% had delays in both areas. An approach to defining proficient and non-proficient readers that allows for the establishment of 3rd grade proficiency based on word reading and comprehension separately may better select early reading measures and criteria that are predictive of reading delay in specific areas of reading. Given my goal of predicting reading proficiency (or disability) in 3rd grade and interest in identifying intervention measures able to predict group membership for late-emerging poor readers and persistent poor readers, I elected to form separate groups of proficient and RD 3rd grade readers based on individual areas of reading.

When forming groups of proficient and RD readers based on single reading skills, previous researchers (e.g., Catts et al., 2011; Vellutino et al., 2008) have relied on measures of word reading and comprehension to establish proficiency status. Absent from these studies are measures of reading fluency, vocabulary, and spelling, each of which add to a comprehensive picture of reading development. Because I planned to explore separate prediction models for different, yet related facets of reading, it was necessary to determine where reading fluency, vocabulary knowledge, and spelling skill fit in. Theoretical and statistical evidence contributed to the placement of each skill within a specific construct.

Reading fluency. Reading fluency refers to a child's ability to read connected text quickly and accurately. In many models of reading development, a child's ability to read lists of words fluently (i.e. word reading fluency) and her ability to read connected text fluently (i.e. passage reading fluency) are associated with one another and with the

ultimate goal of reading: reading comprehension (Hudson, Torgesen, Lane, & Turner, 2012; Schwanenflugel, Meisinger, Wisenbaker, Kuhn, Strauss, & Morris, 2006). Some models posit a direct effect of word reading fluency on text reading fluency (Schwanenflugel et al., 2006) and others posit a reciprocal relationship. Regardless of whether word reading fluency and text reading fluency are causally or reciprocally related, evidence suggests these skills are strongly associated with each other and with reading comprehension.

Word identification tasks differ from word reading fluency tasks in that the first are essentially untimed and the second are timed. In general, fluency tasks require automatic retrieval of words to obtain a high score; conversely, students may obtain high scores on untimed tasks even when they must consciously process individual letters and letter patterns to read words on a list. It is important to note, however, that some word identification measures such as the WRMT-WID subtest allow a maximum of 5 seconds to decode each word from a graded word list and therefore are not completely time-free tasks. Therefore, some degree of automaticity may be required to obtain high scores.

In recent work, Morris and colleagues (2012) tested two competing models of the relationship between word reading, oral reading accuracy (i.e. accuracy of reading words in connected text), oral reading rate, and oral and silent passage reading fluency. In the first model (Figure 1), untimed word reading was posited to contribute directly to oral reading accuracy, which contributed to timed word reading and spelling ability. The latter two skills were posited to predict oral reading rate, which predicted silent reading rate. The alternative model (Figure 2) posited a direct effect of untimed word reading on

timed word reading and spelling, which then contributed to oral reading accuracy. Oral reading accuracy was hypothesized to predict oral reading rate, which was posited to predict silent reading rate.

The models were tested on data collected from 274 students in 2nd through 6th grade who represented a range of reading ability. Results indicated that both models fit the data well. Although Morris et al. were concerned with whether automatic word reading skills preceded or followed oral reading accuracy skills in predicting oral reading rate, the tested models reveal information about relations between untimed word reading, timed word reading, and oral reading accuracy. Across the two models, paths from untimed word reading to oral reading accuracy (Model 1) and timed word reading (Model 2) were strong and significant for students across grades (3rd grade standardized coefficients were B=0.99, p<.05 for each path). For Model 1, the path from oral reading accuracy to timed word reading was also strong and significant for 3rd graders (B=0.93, p<.05). Since the variables explained by untimed word reading skill included oral reading accuracy (a component of passage fluency) and timed word reading (a common predictor of passage fluency), each skill might be subsumed under a similar construct of reading skill. This evidence paired with exploratory factor analysis of the present data (see below) supported the combination of word reading and fluency measures on a single construct of word reading/fluency skill for 3rd graders in this study.

Spelling. The notion that spelling skill and word reading skill are interrelated is not new. Ehri's Theory of Word Reading posits that readers learn sight words by forming connections between word spellings, sounds, and pronunciations (Ehri, 1998).

The combinations of letters that make up each word (i.e. its spelling) helps readers create a semantic map they then use to read known words and decode new words with recognizable patterns more efficiently. Decoding and spelling (or encoding) skills are theorized to work reciprocally to enhance word reading skills (Ehri, 1998; 2005).

Given theoretical and empirical evidence of reading-spelling relations, improving deficit skills in one area should lead to improved performance in the other. To explore whether encoding interventions resulted in improved reading skills for young students, Wieser and Mathes (2011) conducted a best evidence synthesis of intervention studies that included students in kindergarten through third grade and older students reading at or below a third grade level. They included studies that explored the impact of interventions in which spelling instruction was a component on young students' reading performance. Four studies described in the results provided evidence for long-term impact of encoding instruction on at-risk students' word reading, reading fluency, and comprehension performances. Pooled effect sizes for treated groups over controls were 0.7, 0.7, and 0.66 respectively. Also, results of several studies suggested that encoding instruction impacted decoding skill even in interventions where decoding instruction was not a component (see Weiser & Mathes, 2011 for the full review). Based on this evidence, it seemed likely that spelling skill would be associated with word reading, fluency, and/or comprehension skill.

Vocabulary. Vocabulary knowledge has been shown to influence children's reading skill acquisition in early grades (Hart & Risley, 1995) as well as reading comprehension in later grades (Chall, Jacobs, & Baldwin, 1990). Research has also

supported the notion of a reciprocal relationship between reading comprehension and vocabulary development. In a recent study of relations between word reading, comprehension, and vocabulary for good and poor comprehenders aged 8 to 16 (Cain & Oakhill, 2011) reading comprehension skill at age 8 accounted for unique variance in receptive vocabulary growth at age 11, 14 and 16, controlling for previous vocabulary and cognitive skill. Additionally, children with specific difficulties in reading comprehension evidenced slower vocabulary growth than peers with good comprehension skills.

A study of relations between word reading, comprehension, and vocabulary development in a sample of 2,790 Dutch children followed from 1st to 6th grade drew similar conclusions (Verhoeven et al., 2011). Using a cross-lagged design to capture relations between reading skills, the researchers found significant reciprocal relations between reading comprehension and vocabulary development in 2nd – 4th grades. Specifically, vocabulary knowledge in early second grade accounted for about 36% of the variance in 3rd grade reading comprehension scores. The amount of variance in reading comprehension for 4th graders dropped to below 10% when predicted by third grade vocabulary knowledge, though the prediction coefficient was still significant. Associations between vocabulary and word reading were significant, but much smaller especially after accounting for relations with reading comprehension.

Experimental studies investigating the impact of instruction on vocabulary and comprehension performance have demonstrated relations between vocabulary and reading comprehension for older elementary students (Simmons et al., 2010). For their

study, Simmons and colleagues randomly assigned 903 fourth-grade students in 15 schools to a vocabulary-focused, comprehension-focused, or typical instructional conditions for 18 weeks. Results indicated improved content-related reading comprehension for students in both instructional conditions and improved vocabulary knowledge for students in the vocabulary condition (Simmons et al., 2010). Collectively, these studies provide support for meaningful relations between comprehension and receptive vocabulary for students in 3rd grade and beyond.

Exploratory factor analysis. Exploring theoretical evidence linking reading skills was a first step in determining how reading skills might cluster for students in this sample. To better understand the pattern of relationships underlying 3rd grade reading outcomes and identify reading measures that represented similar constructs in this sample, I performed an exploratory factor analysis using raw scores from all 3rd grade measures proposed to contribute to the outcome composites: ORF; the WID, WATT, PC, and WC subtests of the WRMT; PPVT; and TWS. I used the Maximum Likelihood method of estimation and the eigenvalue greater than 1 criterion to determine the appropriate number of factors to describe the data. These decisions resulted in a one factor solution that accounted for approximately 59% of the variance in 3rd grade outcomes. The eigenvalue associated with a second factor was 0.9; this factor accounted for an additional 14% of the variance. Therefore I selected the two factor solution, which accounted for approximately 73% of the variance in reading outcomes. Since the constructs of word reading, fluency, vocabulary, and comprehension are typically correlated, oblique rotation was used to enhance interpretability of results. As can be

seen in Table 5, variables associated with word reading and reading fluency loaded strongly on the first factor (all above 0.7) and variables associated with reading comprehension and vocabulary knowledge loaded strongly on the second factor. The association between factors was moderate (r=0.67). The pattern and strength of relationships between outcome variables were replicated in a factor analysis with Cohort B (Table 5). Based on theoretical and statistical results, two composite variables were created and used to form groups of proficient and non-proficient readers within each cohort: (a) Word Reading/Fluency (WR-F), and (b) Comprehension/Vocabulary (C-V).

Creating RD and proficient reader groups. Outcomes collected at the start of 3rd grade were used to create WR-F and C-V composite measures so that intervention services received in 3rd grade could not influence a student's designation as a proficient or RD reader. In this way, analyses more clearly assess the ability of 1st grade and 2nd grade measures to predict reading proficiency only for students who received access to intervention in those grades. Furthermore, Vellutino et al. (2008) noted that outcomes collected after the summer between grades are more accurate reflections of proficiency or risk since students are required to maintain gains over summer with little to no instruction (also see O'Connor et al., 2005; Vellutino et al., 1996).

The composite variable used to form proficient and RD reader groups according to WR-F outcomes was created as follows: each of the 3rd grade fall ORF, WRMT WID and WATT raw scores, and Spelling scores were converted into Z scores using sample-based means and standard deviations. The Z scores for each measure were combined to form a composite score for WR-F outcomes. Finally, the composite WR-F score was

converted into a sample-based Z score. The resulting variable reflects a student's WR-F performance in the fall of 3rd grade in standard deviation units relative same-grade peers attending similar schools and receiving the same access to equivalent Tier II intervention. These steps were repeated to form proficient and RD groups within Cohort B.

The procedure used to create the C-V composite variable was identical to the procedure used to create the WR-F variable except for the selection of measures. Fall of 3rd grade raw scores for WRMT WC and PC and for the PPVT were used to create the C-V composite. The resulting variable reflects a student's C-V performance in fall of 3rd grade in standard deviation units relative same-grade peers attending similar schools and receiving the same access to equivalent Tier II intervention. These steps were repeated to create proficient and RD groups within Cohort B.

Because the intent of this study was not to determine where the cut-off should be for identifying proficient and RD readers, two distinct cut-off scores below which students would be considered RD were explored in separate analyses using data from Cohort A. For each cut-off score, cases falling above the cut-off plus standard error of the mean for the Z scored composite variable used to form RD groups (*SE*=.07) were removed. The intent was to further segregate the RD and proficient reader groups based on cut-off scores.

The first cut-off considered as a potential indicator of reading disability was a score of 1.5 standard deviation below local normative means; that is, for each of the WR-F and C-V constructs, students with Z scores less than or equal to -1.5 on the composite were identified as RD and those scoring -1.43 and higher were identified as proficient

readers. Setting the RD cut-off at 1.5 standard deviations below the sample mean identifies those students at-risk for severe reading disability. This method of defining groups will be referred to as "Method 1" henceforth. The second method, i.e. "Method 2," imposes a cut-off of 1.0 standard deviation below the local normative mean for defining RD and proficient reader groups (RD $Z \leq -1.0$; proficient $Z \geq -.93$). The $Z<1.0$ cut-off is commonly used in studies where students receiving intervention are divided into good and poor responders (Fuchs et al., 2004; 2008) or RD and non-RD. Overall no more than 2 cases were removed in creating buffers between RD and proficient reader groups.

Tables 7 and 8 display group scores for relevant reading measures, the vocabulary measure, and the CST by proficient and RD reader groups under each of the 1.5 SD and 1.0 SD cut-offs. In selecting the most appropriate method for defining proficient and RD groups for each construct, I explored the following group features: group size and prevalence of proficient and RD readers in the sample, group score differences on construct-relevant tests (i.e. decoding and fluency measures for WR-F construct); group score differences on construct-secondary tests (i.e. C-V measures for the WR-F construct); and substantive meaning of group differences on relevant measures. It is important to recognize that the prevalence estimates provided in Tables 7 and 8 for each of the WR-F and C-V constructs are not mutually exclusive. Students identified as RD under WR-F may also be represented as RD under C-V. Table 6 displays the prevalence rates for RD readers under mutually exclusive categories for WR-F, C-V, and both

constructs. This table may be the more appropriate reference when exploring whether prevalence rates for RD readers fall within an acceptable range.

Group size is a concern given the proposed analyses. Although unbalanced groups do not necessarily negate the use of logistic regression, Spicer (2005) recommends group sizes of at least 20 individuals per group for trustworthy results. Furthermore, models with multiple predictor variables are more stable with larger group sizes. According to Table 7, Method 1 identified 7 RD readers (prevalence of 3.2%) defined by the WR-F construct and 15 RD readers (prevalence of 6.8%) defined by the C-V construct. Under Method 2 (Table 8), 29 RD readers (prevalence of 13.2%) were identified under the WR-F construct and 31 RD readers (prevalence of 14.2%) were identified under the C-V construct. As mentioned, prevalence rates in Tables 7 and 8 are not mutually exclusive. Table 6 shows within RD group and within sample prevalence rates of RD readers with WR-F only, C-V only, or combined delays under Method 1 and Method 2. Sample prevalence rates under Method 1 for WR-F, C-V, and combined delays were 0.9%, 4.6%, and 2.3% respectively. Under Method 1, approximately 8% of the sample was identified as RD readers under at least one of the two constructs. Conversely, under Method 2, prevalence rates for RD readers with WR-F only, C-V only, or combined delays were 6.4%, 7.3%, and 6.8% respectively. Approximately 21% of the sample was identified as RD readers under at least one of the two constructs under Method 2.

Multiple Analysis of Variance (MANOVA) was executed to investigate differences in 3rd grade outcomes between proficient and RD reader groups under each

method of identification and for each of the WR-F and C-V constructs. Tables 7 and 8 shows mean scores, between group difference scores, *p*-values, and Effect Sizes (*ES*) for differences between proficient reader and RD groups on individual outcome measures. The tables are organized by cut-off method (i.e. Method 1 or Method 2) and by reading construct (i.e. WR-F or C-V). Cohen's *F* effect size is provided as an unbiased alternative to the upwardly biased Eta-Squared (Cohen, 1988). Cohen's *F* is calculated by taking the square root of the ratio of Eta-squared to 1-Eta Squared and is measured in standard deviation units. Effects are interpreted as small (.10), medium (.25) or large (.40) according to Cohen (1988).

Under Method 1 (Table 7) and for each of the WR-F and C-V constructs, differences between proficient and RD readers were significant (all $ps \leq .001$) for construct-relevant measures, with effect sizes ranging from .24 to .44. Differences for construct-secondary measures were also significant in every case with *ESs* ranging from .24 to .44, except for PPVT on the WR-F construct (*ES*=.1). Given the association between word reading, fluency and reading comprehension skill for elementary-aged students, it was not surprising that students who struggled in one area (e.g. comprehension) commonly struggled in the other.

Group differences on construct-relevant measures under the WR-F construct were substantively meaningful. For example, the average ORF score for RD students fell well within the at-risk range (at-risk ≤ 52wcpm) for 3rd grade fall according to DIBELS benchmarks (*M*=24.71, *SE*=8.8). Conversely, the average ORF for proficient readers (*M*=78.02. *SE*=1.7) fell near the no-risk range (No Risk ≥ 77wcpm). A similar trend was

found for WRMT WID, WATT, and Spelling scores. Differences on construct-relevant outcomes for the C-V construct showed a similar pattern of results as those for the WR-F construct. Namely, proficient readers outscored RD readers on PPVT, with an *ES* of .43. Proficient readers also outscored RD readers on WRMT PC and WC by approximately half-standard deviation each.

Third grade CST scores were used to help interpret proficient and RD readers' performance relative to state-standards. According to Table 7, average CST scores for RD readers under the WR-F construct (*M*=261.43, *SE*=17.9) fell within the Below Basic/Far Below Basic range and scores for proficient readers (*M*= 321.97, *SE*=3.4) fell within the Basic Range, which was also the whole-sample average range (see Procedures and Table 3). Under C-V, CST scores were significantly higher for the proficient group (*M*=324.68, *SE*=3.3) compared to the RD group (*M*=255.57, *SE*=12.1). Like with WR-F outcomes, RD students scored in the Far Below Basic/Below Basic range and proficient students in the Basic range.

Table 8 outlines differences in outcomes between proficient reader and RD groups for each of the WR-F and C-V constructs under Method 2. Differences between groups within each of the WR-F and C-V constructs were significant on all outcomes (*p*s< .01). *ES'*s were large and differences on construct-relevant measures were substantively meaningful. *ES*'s on construct-relevant measures were also generally larger than those on construct-secondary measures under each construct. Under WR-F for example, effect sizes for ORF, WID, WATT, and Spelling ranged from .46 to .56, while effect sizes for Word Comprehension and PPVT were .39 and .18 respectively. The

effect size for PC was large (*ES*=.56), likely since WRMT PC subtest also relies on the ability to read words. Proficient readers under WR-F read 80 wcpm on average, which falls in the "low-risk" range according to DIBELS benchmarks. Conversely, RD readers under WR-F read approximately 46 wcpm on average, which falls within the "at-risk" range according to DIBELS benchmarks. Similar trends in effect size differences between proficient and RD readers occurred under the C-V construct. Effect sizes for PC, WC, and PPVT ranged from .45 to .61, while effect sizes for ORF, WID, WATT, and Spelling were smaller (range= .25 to .45). Differences between groups on 3rd grade CSTs were also significant (*p*s <.001) and substantively meaningful under each of the WR-F and C-V constructs, with effect sizes of .49 and .39 respectively. In each case, proficient readers scored in the Basic range and RD readers scored in the Far Below Basic/Below Basic range.

The above analysis of differences in 3rd grade outcomes between proficient and RD readers for WR-F and C-V constructs under Methods 1 and 2 suggests that under each Method and for each construct, groups differed in meaningful ways. The patterns of differences between groups across Methods and constructs were similar. In all cases, proficient readers had higher scores than RD readers. In most cases, effect sizes for differences between groups were larger for construct-relevant measures compared to construct-secondary measures. Prevalence rates for proficient and RD groups under each construct and Method are acceptable for the purposes of this study. When Method 2 was repeated with Cohort B data, results mirrored those from Cohort A (see Table 9). Therefore, to ensure greater power of estimation and model stability, I focused on

Method 2 to distinguish between groups for the WR-F and C-V outcomes. I make note of how outcomes might differ under Method 1 definition of RD and proficient groups as needed.

Data Analysis

Missing data. Missing data is a concern in longitudinal studies where attrition and sporadic non-response are common. Data for this study were derived from an existing database given the condition that participants were present and had *access* to intervention during grades 1 through 3. Two-hundred thirty seven cases met criteria for inclusion for Cohort A. Of these 237 cases, 10 were removed due to missing data on all 3rd grade outcomes. An additional 8 were removed due to missing data on more than half of the predictor variables. Ninety-five percent of the remaining cases for Cohort A had complete data on 3rd grade outcome measures. Of the remaining 5%, 9 cases (4.1% of the sample) had missing values on 3rd grade PPVT and 1 (0.4%) had missing values on 3rd grade Spelling; 1 case (0.4%) was missing outcome data for both measures. Scores for same measures given at the previous time point were used to replace missing values for 2 of 9 cases on PPVT and both cases on Spelling. Mean substitution was used to replace the remaining 7 missing values.

One hundred eighty-one cases from Cohort B met inclusion criteria. Of these 181 cases, 13 were removed due to missing data on all outcome variables. Ninety-four percent of remaining cases had complete data on 3rd grade outcomes. Of the remaining 6%, 4 cases (2.4% of the sample) were missing values on PPVT, 5 cases (3%) were missing data on fall Spelling, and 1 case (0.6%) was missing data on the WRMT subtests

and on ORF. Scores for same measures given at nearby time points were used to replace missing values for 2 of the 4 cases on PPVT, all cases on Spelling, and the case with missing data on WRMT and ORF. Mean substitution used to replace the remaining 2 missing PPVT values.

For each cohort, no more than 5% of data were missing for any predictor variable. Missing values on predictor variables were handled using the Maximum Likelihood Robust (MLR) method of estimation for logistic regression analyses. Maximum Likelihood is a useful method of estimation for models with missing data because unlike deletion methods (e.g. listwise, pairwise deletion), ML uses all available data to derive unbiased parameter estimates (Allison, 2002).

Logistic regression analyses. To address research question 1, a preliminary analysis to identify 1st and 2nd grade reading measures that adequately predict membership in RD or proficient reader groups by 3rd grade was required. Logistic regression analysis was selected as the analytical framework to accomplish this task, and was conducted using data from Cohort A. Two separate logistic regression analyses were executed in Mplus using the MLR method of estimation. Proficient and RD 3rd grade readers were categorized into groups based on composite scores for WR-F and C-V measures separately. Under each of the WR-F and C-V constructs, RD readers were those whose composite reading score fell 1.0 standard deviation below the sample mean (Method 2; see Method section). In a series of logistic regression analyses, 1st and 2nd grade reading measures were entered stepwise to predict group membership and the most influential measures were included in further analyses. Demographic covariates were

entered with 1st grade predictors in Step1 and retained if significant. Although the purpose of this analysis was to identify significant predictors of group membership and not to identify superior models for identification, AIC, BIC, and the negative 2 x loglikelihood fit statistics are provided.

Classification analyses. After identifying significant 1st and 2nd grade predictors of RD group membership using logistic regression and the definition of RD under Method 2, significant predictors were paired with a-priori selected responsiveness criteria to identify measure/criteria combinations most effective in the prediction of 3rd grade RD. This analysis served as a direct response to research question 1. Two by two contingency tables were used to calculate sensitivity and specificity of predictive measures in identifying children who would be designated as proficient or RD readers by 3rd grade. Sensitivity and specificity of measure/criteria combinations were evaluated against each other and the field standard of .9 for sensitivity and .8 for specificity (Jenkins et al., 2003). In the current context, sensitivity describes the power of 1st and 2nd grade reading measures to detect those children who will be classified as RD according to composite scores on measures administered in 3rd grade. Correct classification occurs when students are identified as poor responders on 1st and 2nd grade measures and as RD on 3rd grade outcomes. Specificity describes the power of the 1st and 2nd grade reading measures to identify those children who will be classified as proficient readers according to composite scores on measures administered in 3rd grade. For specificity, correct classification occurs when children are designated as good responders according to 1st and 2nd grade reading measures and as proficient readers based on 3rd grade outcomes.

Criteria used to designate students as good or poor responders varied depending on the type of reading measure. ORF was paired with Final Benchmark/Final Status, Low Growth, DD, and percentile cut-point criteria, while WIF and standardized measures were paired with percentile cut-point criteria. Growth estimates for the ORF Low Growth criterion were generated in Mplus using the Structural Equation Modeling framework (Raykov & Marcoulides, 2008).

Replication. Research question 2 required exploration of the extent to which results gained from logistic regression and classification analyses with Cohort A replicated with a second cohort of students (i.e. Cohort B). As noted, factor analyses of 3rd grade outcomes for Cohort B supported a 2 factor solution with loadings on a WR-F construct and C-V construct similar to those of Cohort A (Table 5). Groups of RD and proficient readers under WR-F and C-V constructs were formed using the same procedures to form groups in Cohort A. Two cases were removed in creating the buffer between RD and proficient reader groups for the WR-F construct, and zero cases were removed in creating the buffer for the C-V construct.

To determine the extent to which results of the logistic regression replicated, the most parsimonious model that effectively predicted RD group membership for Cohort A was imposed on scores for Cohort B and sensitivity and specificity indices were compared between cohorts. Finally, measure/criteria combinations explored in Cohort A were repeated with Cohort B and sensitivity and specificity indices were compared.

Chapter 4: Results

All data were screened for scoring and input errors, and for univariate and multivariate outliers. Univariate outliers were detected using box plots generated in SPSS and multivariate outliers were identified through visual inspection of the data. Errors and outliers were reconciled with scores from original data sources**.**

To determine whether significant variance in WR-F and C-V outcomes was present between classrooms, a two-level unconditional model with students at Level 1 and classrooms at Level 2 was executed for each outcome within each cohort using the HLM 7 software (Raudenbush, Byrk, Cheong, Congdon, & Toit, 2011). Intraclass Correlation (ICC) was calculated to identify the percentage of variance in the outcome that existed between Level 2 units (Raudenbush & Byrk, 2002). The outcomes used for these analyses were composite scores on each of the WR-F and C-V constructs. Within Cohort A, the ICC indicated that 2% of variance in WR-F composite scores was between classrooms, χ^2 (20) = 29.25, p = .08. Alternatively, 3% of variance in C-V composite scores was between classrooms, χ^2(20) = 35.57, p = .02. Although significant variance on C-V outcomes was between classrooms, the proportion of variance was negligible. Identical models were repeated to determine whether significant variance in WR-F and C-V outcomes was present between classrooms for students in Cohort B. For the WR-F outcomes model, the ICC indicated that 3.4% of variance in outcomes was between classrooms, χ^2 (18) = 25.26, p = .147. For the C-V outcomes model, the ICC indicated that 1.8% of variance in outcomes was between classrooms, χ^2 (18) = 20.29, p = .316. Variance between classrooms was minute for each model estimated, with no variance

exceeding 3.5%. Additionally, in all but one case (i.e. the C-V model for Cohort A), the chi-square test of the null that variance at Level 2 is equal to zero was non-significant. Therefore, data were analyzed without estimating classroom effects on outcomes.

Means and standard deviations are provided for demographic and assessment variables for each Cohort in Table 3. Non-parametric tests of distributional differences in gender, ethnicity, and English learner status between cohorts were conducted in SPSS 18. Mann-Whitney U tests revealed no significant differences in the distributions for gender, dummy variables for Hispanic and African American, or English learner status between Cohorts, all *p*s> .12. Average 1st grade CELDT scores were equivalent between cohorts, $t(201) = -.175$, $p = .862$.

Table 3 provides means and standard deviations for 3rd grade outcomes and for 1st and 2nd grade predictors for each cohort. *P*-values associated with univariate test of differences between cohorts are provided where significant. The significance level was adjusted to $\alpha = .003$ to correct for the 16 comparisons between groups. Significant differences were evident for the 2nd grade WATT and PC subtests for WRMT, where scores for Cohort B were higher on average than those of Cohort A.

Correlations among 3rd grade outcomes are provided for each cohort in Table 10; correlations among predictors and between predictors and outcomes are provided for each cohort in Table 11. Note that correlations for Cohort A are reflected on the rows and below the diagonal and those for Cohort B are on the columns and above the diagonal in Table 10 and in the top section of Table 11. Correlations among WR-F

measures and C-V measures were moderate to strong and significant. Overall, correlations for Cohort B were stronger than those for Cohort A.

Cohort A

A MANOVA was executed in SPSS 18 to determine whether scores on outcome composites and predictor variables varied by gender, ethnicity, or English learner status. Box's M test warranted acceptance of the assumption of homogeneity of covariance matrices at α=.01, $F = 1.176$, $p > .01$. Applying a more conservative significance level is suggested when interpreting results of Box's *M* due its notorious sensitivity to non-normality (Raykov & Marcoulides, 2008). Levene's test of equal error variances for univariate analyses revealed no significant differences and therefore no univariate homogeneity assumption violations.

The MANOVA of all variables by gender was significant, Wilks's $\Lambda = .861$, $p = .001$. Univariate tests revealed significant differences on the composite score for WR-F, $F(1, 198) = 3.79$, $p = .05$, and V-C, $F(1,198) = 5.68$, $p = .013$. In each case, males outscored females. Therefore, gender was entered as a covariate in each logistic regression analysis.

The MANOVA of all variables by English Learner (EL) status was non-significant Wilks's $\Lambda = .953$, $p = .496$. Univariate tests revealed no significant differences on any variables at a significance level of .05. The univariate test for 1st grade TOLD was significant at $\alpha = .10$, $F(1,198) = 2.8$, $p = .09$, indicating lower scores for EL students. To eliminate possible impact of EL status on outcomes, EL status was entered as a covariate in analyses involving C-V outcomes.

The MANOVA of all variables by ethnicity was non-significant, Wilks's Λ Hispanic = .923, p = .113, and Wilks's Λ African American = .960, p = .641. Univariate tests revealed no significant differences on any measure for Hispanic or African American participants at the .05 significance level. The univariate test for the C-V composite was significant for Hispanic at α=.10, $F(1,198) = 15.89$, p = .07, indicating lower scores for Hispanic students. To eliminate possible impact of ethnicity on outcomes, a dummy-coded variable for Hispanic was entered as a covariate in analyses involving C-V outcomes.

Logistic regression 1: word reading/fluency. Logistic regression was used to identify 1st and 2nd grade reading measures that were most useful in the prediction of reading proficiency/disability in 3rd grade. Recall that students were identified as RD under Method 2 if their scores on the composite WR-F construct (comprising 3rd grade ORF, WID, WATT, and Spelling measures) fell 1.0 standard deviation or more below the sample mean. Predictors were entered stepwise, with 1st grade measures entered in the first step and if significant, included in Step 2 with 2nd grade measures. Demographic covariates were entered in the first step and retained in the second step if significant. Gender was coded 0 = female, 1 = male; Hispanic was coded 0 = all other, 1 = Hispanic. The final model reflects only significant predictors from Step 2.

Assumptions for logistic regression are less restrictive than those that apply to ordinary least squares regression (Spicer, 2005). Independence of cases and absence of multicollinearity should be investigated though, since violations of these assumptions may impact outcomes. Predictors in this study had correlations that ranged from

moderate (.4) to high (>.8). To determine if multicollinearity posed a serious issue, I calculated the Variance Inflation Factor (VIF) for each highly correlated variable. VIFs were compared to a maximum acceptable value of 10 (Cohen, Cohen, West, & Aiken, 2003); no VIFs for any potentially problematic variables exceeded 2.7. Therefore, multicollinearity was not considered to significantly impact model results.

Since participants were nested within classrooms, the assumption of independence may have been violated. However, HLM analyses revealed negligible variance in outcomes between classrooms. No other violations of independence were apparent. Therefore, this assumption was considered fulfilled.

Logistic regression coefficients, standard errors, *p*-values, and odds ratios are provided in Table 12. Gender was significant in Step 1, b= -1.86, p = .001, and Step 2, b = -2.36, p = .008, reflecting overrepresentation of females in the RD group for this sample. Thus, gender was retained as a control variable for remaining analyses. The coefficient for 1st grade WIF was significant in Step 1, b = -.071, p = .032, but was no longer significant once 2nd grade predictors were entered in Step 2, b = -.001, p = .982. Of the 2nd grade predictors entered at Step 2, ORF and WRMT PC were significant, b = -.11, p <.001; b = -.16, p <.001, respectively. The final model included only significant predictors from Step 2: gender, b = -2.30, p = .007; 2nd grade ORF, b = -.10, p < .001; and 2nd grade Passage Comprehension, b = -.21, p < .001.

Beta coefficients for each predictor express the change in log odds of being in the group coded 1 (i.e. the RD group) for every 1 unit change in the predictor. All significant coefficients were in the expected direction; that is, for each measure, unit increases in the

predictors were associated with reduced log odds that a case would belong to the RD group. Exponentiation of log odds eases interpretation; taking 1 minus the log odds and multiplying by 100 provides the percent increase (for positive values) or decrease (for negative values) in odds of being in the RD group for every 1 unit increase on the predictor variable. For example from Table 12, 1st grade WIF in Step 1 has an odds ratio of .93. Subtracting 1 from this value and multiplying by 100 shows that for every 1 unit increase on WIF, the odds of being in the RD group decreased by 6.8%.

The purpose of this logistic regression analysis was to identify 1st and 2nd grade measures best able to distinguish between RD and proficient reader groups, so tests of model deterioration or improvement are not of primary relevance. Nevertheless, Table 12 indicates lower AIC and BIC values at each step compared to the Null model, indicating superior model fit for each of the conditional models compared to the Null (Raykov & Marcoulides, 2008). Traditional Chi-Square statistics are not provided in Mplus. Therefore, Chi-square difference tests of each step compared to the Null model were conducted using changes in negative loglikelihood values with degrees of freedom equal to the number of independent variables in the model[1]. The change in negative loglikelihood of the Null compared to Step 1 (loglikelihood null = 5489.4, $df = 0$; Step 1 = 1826.37, $df = 4$) was significant, indicating significant improvement in the model with the addition of predictors, $\chi^2(4) = 3663.03$, $p < .001$. The change in negative

[1] Chi Square statistics are not provided in Mplus; therefore, difference tests were conducted using the -2 log likelihood for the intercept-only model compared to prediction models with degrees of freedom equal to the number of independent variables. The null hypothesis states that addition of the predictors does not add to knowledge of group membership. Rejection of the null indicates model improvement. This test is analogous to a test of R-square change in step-wise regression analysis and is an appropriate alternative index of model fit (Spicer, 2005)

loglikelihood from the null to Step 2 was also significant, $\chi^2(6) = 1407$, $p < .001$, as was the final model from each of the Null, $\chi^2(3) = 3702.34$, $p < .001$, Step 1, $\chi^2(1) = 39.31$, $p < .001$, and Step 2, $\chi^2(3) = 2232.23$, $p < .001$.

Logistic regression 2: vocabulary/comprehension. As with word WR-F outcomes, logistic regression was used to identify 1st and 2nd grade reading measures that were most useful in the prediction of RD in V-C in 3rd grade. Recall that students were identified as RD under Method 2 if their scores on the composite V-C construct (comprising 3rd grade PPVT, WC, and PC measures) fell 1.0 standard deviation or more below the sample mean. Predictors were entered stepwise, with 1st grade measures entered in the first step and if significant, included in Step 2 with 2nd grade measures. Demographic covariates were entered in the first step and retained in the second step if significant. Gender and Hispanic retained their coding schemes; 1st grade English Learner status was coded 0=English Only, 1=English Learner. The final model reflects only significant predictors from Step 2.

Logistic regression coefficients, standard errors, *p*-values, and odds ratios are provided in Table 13 Gender was significant in Step 1, $b = -1.09$, $p = .0$, and Step 2, $b = -1.12$, $p = .02$, reflecting overrepresentation of females in the RD group for this sample. Thus, gender was retained as a control variable for remaining analyses. First grade TOLD was the only significant predictor of RD status in Step 1, $b = -.17$, $p = .002$, but was no longer significant with the inclusion of 2nd grade predictors, $b = -.06$, $p = .48$. Significant 2nd grade predictors were standard scores on TOLD, $b = -.29$, $p = .01$, and

ORF raw scores, b = -.02, p = .03. The final model contained gender, 2nd grade TOLD, and 2nd grade ORF.

As with the WR-F model, significant coefficients were in the expected direction. Additionally, AIC and BIC values favored the conditional models over the Null and the final model over Steps 1 and 2. Chi-square change statistics indicated significant model improvement at each step and the final model compared to the Null.

Differences with Method 1. Based on considerations outlined in the Method section, I chose to describe proficient and RD groups based on composite measure Z scores relative to a 1.0 standard deviation below the sample-mean cut-off. The choice of cut-off to define proficient and RD groups is not the focus of this dissertation; however, comparison of logistic regression models based on Method 2 and Method 1 cut-offs is illustrative.

For the WR-F model under Method 1, coefficients on all demographic variables were non-significant. WIF was a significant 1st grade predictor, b = -.369, p = .003, and was of greater magnitude than under Method 2. Once 2nd grade predictors were entered, WIF was no longer significant. Second grade ORF was the only significant 2nd grade predictor, b = -.377, p = .04, and was of greater magnitude than under Method 2. Although Passage Comprehension was a significant predictor of RD under Method 2, it was non-significant under Method 1. The final model under Method 1 consisted of gender and 2nd grade ORF.

For the V-C model under Method 1, the only significant demographic variable was gender, b = -.952, p = .02. Similar to results from Method 2, TOLD was the only

significant 1st grade predictor of RD, b = -.23, p = .008, and was of larger magnitude. After accounting for gender and 1st grade TOLD, no 2nd grade variables were significant at α = .05. However, in a model with gender and 2nd grade predictors (omitting 1st grade TOLD), the coefficient on 2nd grade TOLD was significant, b = -.213, p = .034, and of similar magnitude as under Method 2. A final model with the same predictors under Method 2 was explored, regressing RD on gender, 2nd grade TOLD and 2nd grade ORF. Results were similar to the final model under Method 2; effects for each of gender, 2nd grade TOLD and 2nd grade ORF were significant and of similar magnitude as those resulting from analyses under Method 2.

Classification analyses. To determine whether any single significant predictor of RD from the logistic regression analyses adequately identified proficient and RD readers, two-by-two contingency tables were created posing responder status against RD classification. Significant predictors of RD in WR-F were 1st grade WIF and 2nd grade ORF and WRMT PC. Although WIF was no longer significant once 2nd grade measures were added to the model, WIF was included in the classification analysis to explore sensitivity and specificity of RD prediction resulting from a 1st grade reading measure. Sensitivity and specificity of all explored measure/criteria combinations are provided in Table 14.

Since no benchmarks exist for WIF, 1st grade WIF was paired with 25th and 33rd percentile criteria and sensitivity and specificity indices for the prediction of RD were calculated for each measure/criteria combination. Children scoring below the relevant percentile cut-point were designated poor responders. Under the 25th percentile criterion,

75% of children with RD and 81% of proficient readers were correctly classified. The 33rd percentile criterion correctly classified 82.1% and 76.1% of RD and proficient readers, respectively.

Second grade ORF was paired with Final Benchmark, Low Growth, DD, and percentile cut criteria. According to DIBELS benchmarks (Good et al., 2001) children should read approximately 90 wcpm by spring of 2nd grade; therefore, the Final Benchmark criterion was set at 90 wcpm. Children reading 90 or more wcpm were categorized as good responders and those reading fewer than 90 wcpm were categorized as poor responders (i.e. possibly RD). According to the classification table (Table 14), 100% of children with RD were correctly classified under the Final Benchmark criterion and 55.9% of proficient children were correctly classified. To improve specificity, the Final Benchmark criterion was reduced to the highest score in the RD group: 75wcpm. Under the new criterion, sensitivity remained at 100%[2] and specificity improved to 78.8%.

The ORF Low Growth criterion was specified such that children with growth ≤ 1 standard deviation below the sample average growth were classified as poor responders. This criterion produced very poor sensitivity rates and adequate specificity, correctly classifying 37.9% of children with RD and 89.4% of proficient readers. Children were identified as poor responders under DD in ORF if their spring scores fell below the final benchmark (90wcpm) and if their growth in ORF fell 1 or more standard deviations

[2] by design, the final benchmark score was lowered to reflect the lowest ORF score that would still yield sensitivity of 100%

below the sample average growth. Under this criterion, 40.7% of children with RD were correctly classified and 95.1% of proficient readers were correctly classified.

When paired with 25th and 33rd percentile-cut criteria, poor responders on ORF were those whose scores fell below the percentile cut-point. The cut-score for the 25th percentile on ORF was 73 wcpm. Under this criterion, sensitivity and specificity were 85.2% and 83.2% respectively. Under the 33rd percentile criteria (ORF $\leq$ 77 wcpm), sensitivity improved to 100% and specificity was reduced to 75.5%. The 33rd percentile criteria correctly classified an additional 4 RD readers and misclassified an additional 14 readers compared to the 25th percentile criterion.

Finally, 2nd grade WRMT Passage Comprehension was paired with 25th and 33rd percentile criteria to form group of good and poor responders. For each criterion, poor responders were those whose Passage Comprehension standard scores fell below the percentile cut-point. The 25th percentile cut-off provided the best balance between sensitivity and specificity rates of the three percentile criteria. Under this criterion, 79.3% of children with RD and 81.6% of proficient readers were correctly classified. Using the 33rd percentile criterion improved sensitivity to approximately 90% and reduced specificity to 73.5%.

Significant predictors for RD in V-C were 2nd grade TOLD and 2nd grade ORF. Although 1st grade TOLD was no longer a significant predictor of group membership after the inclusion of 2nd grade measures, it was retained in classification analyses to determine whether it was useful as an isolated predictor. To identify good and poor responders, 1st and 2nd grade TOLD were paired with 25th and 33rd percentile criteria and

ORF was paired with Final Benchmark, Low Growth, DD, and percentile-cut criteria. Sensitivity and specificity indices for each measure/criteria combination are provided in Table 15.

First grade TOLD measure/criteria combinations yielded poor sensitivity and specificity indices overall. The 25th percentile criterion produced sensitivity of 50% and specificity of 71.8%. The 33rd percentile criterion improved sensitivity to 56.3% at the expense of specificity, which dropped to 66.3%.

Second grade TOLD measure/criteria combinations yielded similarly poor sensitivity and specificity rates at each of the 25th and 33rd percentile cut-offs. The 33rd percentile criterion provided the most adequate balance, yet sensitivity and specificity were inadequate (sensitivity = 72.7%; specificity = 66.9%). The 25th percentile criterion correctly classified 66.7% and 78.7% of RD and proficient readers, respectively.

Sensitivity and specificity associated with ORF varied depending on paired criterion. The ORF Final Benchmark criterion produced a sensitivity rate of 81.3% and specificity rate of 54.2%. Imposing the Low Growth criteria on ORF resulted in poor sensitivity (24.2%) and adequate specificity (87.5%). The DD criterion also yielded a very poor sensitivity rate, correctly classifying 25% of children with RD. Specificity under DD was adequate, at 93.2%.

Percentile criteria paired with ORF produced poor balance between sensitivity and specificity overall. The 25th percentile correctly classified 46.9% of RD readers and 78.2% of proficient readers. The 33rd percentile correctly classified an additional 4

children, improving sensitivity to 59.4%. An additional 14 proficient readers were misclassified, with a sensitivity of 70.4%.

Cohort B

Similar to Cohort A, MANOVAs were executed with Cohort B to determine if scores on predictors or outcomes varied by demographic variables. Box's M test was non-significant at $\alpha = .01$ ($p = .02$) indicating no violation of the homogeneity of variance/covariance assumption. A MANOVA of all variables by gender was non-significant, Wilks's $\Lambda = .947$, $p = .720$; however, univariate tests revealed significant differences on 1st grade WIF, $F(1, 151) = 4.41$, $p = .04$, and 1st and 2nd grade ORF, $F(1,151) = 5.8$, $p = .02$ and $F(1,151) = 4.9$, $p = .03$, respectively. In each case, females outscored males.

A MANOVA of all variables by English Learner (EL) status was significant, Wilks's $\Lambda = .860$, $p = .025$. Univariate tests revealed a no significant differences on any single measure at $\alpha = .05$, and one significant difference favoring native English speakers on the composite variable for C-V at $\alpha = .10$, $F(1,151) = 3.74$, $p = .06$. The significant effect of EL on the C-V outcome is similar to that of Cohort A.

A MANOVA of all variables by Ethnicity was significant for Hispanic Wilks's $\Lambda = .851$, $p = .02$ and non-significant for African American, Wilks's $\Lambda = .926$, $p = .428$. Univariate tests for Hispanic revealed significant differences on 1st grade TOLD, $F(1,151) = 6.6$, $p = .01$ and ORF, $F(1,151) = 4.53$, $p = .04$; 2nd grade ORF, $F(1,151) = 4.0$, $p = .05$, Word Identification, $F(1,151) = 4.64$, $p = .03$, and Passage Comprehension,

$F(1,151) = 5.6$, $p = .02$. Univariate tests for African American revealed no significant differences on any measure.

Overall, differences on outcomes among demographic groups are more similar than different across cohorts, which is expected given samples were collected from students attending the same schools and receiving instruction from the same teachers during approximately the same time periods. The exception is the gender effect. Males outscored females on 3rd grade WR-F and C-V composite measures in Cohort A, and females outscored males on select 1st and 2nd grade predictor variables in Cohort B. Although significant at $\alpha = .05$, no gender differences were significant after significance levels were adjusted for multiple comparisons using Bonferroni correction (i.e. at $\alpha = .004$). Nevertheless, the gender coefficient was significant in WR-F and C-V regression analyses with Cohort A and was therefore a useful covariate.

Replication

Logistic regression. Research question 2 queried the extent to which logistic regression models and classification analyses executed with Cohort A replicated with a new sample. Derivation of Logit and Predicted Probability scores used in logistic model replication proceeded as follows. Scores on relevant measures from Cohort B were entered into the final logistic regression equation for each of the WR-F and C-V constructs to derive Logit terms. Logit terms were used to calculate each case's probability of RD classification using the following formula:

$$\text{Probability} = \frac{1}{1+e^{-(logit)}}$$

For analyses on WR-F outcomes, prior probabilities were set to the proportion of cases with RD in WR-F from Cohort A (13%). Setting prior probabilities to the proportion of cases with RD in WR-F from Cohort B (i.e.14%) resulted in no change in model sensitivity and specificity indices. Therefore, predicted probabilities at or above 13% were classified as RD for Cohorts A and B. The same method for setting prior probabilities was adopted for analysis of C-V outcomes. In this case, priors were set to 14% for each cohort. Previous studies with similar analyses have also used proportions associated with RD groups as the prior probability cut-off (e.g. Compton, D. Fuchs, L. S. Fuchs, & Bryant, 2006); this procedure appears acceptable for making cut-point decisions for priors in logistic regression.

The final logistic regression model used for the prediction of RD in WR-F for Cohort A was:

$$RD = 26.167 - 2.302*Gender - 0.101*ORF\ 2^{nd} - 0.209*PC\ 2^{nd}$$

Model sensitivity and specificity for Cohort A were 86.2% and 88.9%, respectively (Table 16). Three cases with RD were misclassified as proficient and 25 proficient readers were misclassified as RD. Replication of the final model on Cohort B resulted in sensitivity of 80% and specificity of 89.2%. Using the model derived from Cohort A, 5 cases with RD were misclassified as proficient and 15 proficient readers were misclassified as RD.

The final logistic regression model for the prediction of RD in C-V for Cohort A was:

$$RD = 4.576 - 0.923*Gender - 0.038*ORF\ 2^{nd} - 0.408*TOLD\ 2^{nd}$$

Model sensitivity and specificity for Cohort A were 84.4% and 76.7% respectively (Table 16). Five cases with RD were misclassified as proficient and 41 proficient readers were misclassified as RD. Replication of the model on Cohort B resulted in sensitivity of 82.1% and specificity of 69.8%. Using the model derived from Cohort A, 5 cases with RD were misclassified as proficient and 42 proficient readers were misclassified as RD.

Replication of logistic regression models was supplemented by replication of classification analyses using measure/criteria combinations for the prediction of RD in each of WR-F and C-V constructs. Although only the final logistic regression model for each construct was subject to replication, 1st grade measure/criteria combinations were included in the classification replication to identify those effective in the prediction of RD. Predictive 2nd grade measures from logistic regression analyses were also paired with responsiveness criteria in the replication. Sensitivity and specificity of prediction were compared between Cohorts A and B to determine the extent to which results from Cohort A replicated.

Classification analyses. For Cohort B, 1st grade WIF and 2nd grade Passage Comprehension scores were paired with 25th and 33rd percentile criteria to identify groups of good and poor responders to instruction. Second grade ORF was paired with Final Benchmark, Low Growth, DD, and percentile-cut criteria to determine responsiveness. Sensitivity and specificity rates for classification into the WR-F RD group are provided in Table 14.

The 25th percentile criterion for 1st grade WIF provided the best balance between sensitivity and specificity. Four RD readers and 35 proficient readers were misclassified,

resulting in sensitivity of 80.8% and specificity of 81.9%. The 33rd percentile criterion correctly identified an additional 2 RD students, improving sensitivity to 88.5%; eight additional proficient readers were misclassified, reducing specificity to 76.1%.

The 25th percentile criterion for 2nd grade Passage Comprehension provided the best balance between sensitivity and specificity. Five RD and 23 proficient readers were misclassified, resulting in sensitivity of 80% and specificity of 83.6%. An additional 2 RD readers were correctly classified under the 33rd percentile criterion, improving sensitivity to 88%. Thirteen additional proficient readers were misclassified, reducing specificity to 74.3%.

The Final Benchmark, Low Growth, DD, and percentile-cut criteria used to identify good and poor responders according to 2nd grade ORF for Cohort B were identical to those used with Cohort A. Under Final Benchmark (Spring ORF≤ 90 wcpm), sensitivity was 100%. Although no RD readers were misclassified, more than half of the proficient readers were misclassified, resulting in specificity of 43.2%. To improve specificity, the Final Benchmark criterion was lowered to the highest score earned by any RD reader on the spring of 2nd grade ORF measure (i.e. 70 wcpm). This change did not impact sensitivity (by design, the score was reduced to the lowest score that would capture 100% of RD readers) and improved specificity to 69.1%.

The LG criterion (i.e. growth on ORF ≤ 1 standard deviation below the sample mean) resulted in poor sensitivity and excellent specificity. The LG criterion misclassified 10 of 16 RD readers with a sensitivity of 69.6%. Seven proficient readers were misclassified, resulting in specificity of 95%. Case classification for DD was

identical to that for LG. Every case with growth less than or equal to 1 standard deviation below the sample mean also had spring ORF scores lower than 90 wcpm. Thus, sensitivity and specificity were also 69.6% and 95%, respectively.

The 25th percentile criterion on ORF produced a cut-score of 61 wcpm. Under this criterion, 84.6% of RD readers and 85.6% of proficient readers were correctly classified. The cut-score produced by the 33rd percentile criterion was 67 wcpm. This criterion correctly classified an additional 3 RD readers, with a sensitivity of 96.2%. Nine additional proficient readers were misclassified as RD, yielding a specificity of 79.1%.

Significant predictors of RD in C-V were 1st and 2nd grade TOLD and 2nd grade ORF. TOLD was paired with 25thand 33rd percentile criteria and ORF was paired with Final Benchmark, Low Growth, DD, and percentile criteria to identify good and poor responders to instruction. Sensitivity and specificity rates for each measure/criteria combination are provided in Table 15.

The percentile criteria paired with 1st grade TOLD produced poor sensitivity and specificity rates overall. The 25th percentile criterion was inadequate, with sensitivity of 46.4% and specificity of 79.4%. The 33rd percentile criterion improved sensitivity to 71.4% and reduced specificity to 61%.

For 2nd grade TOLD, cut-scores for the 25th and 33rd percentiles were identical. Reducing the 25th/33rd percentile cut-score by 1 unit defined the 22nd percentile. Therefore, the 22nd and 33rd percentiles served as cut-offs for 2nd grade TOLD. Each percentile cut-off produced poor sensitivity, misclassifying no fewer than 8 of 28 RD

cases as proficient. The 22nd percentile resulted in sensitivity of 53.6% and specificity of 83.6%. The 33rd percentile improved sensitivity to 64.3% and reduced specificity to 72.9%.

ORF paired with Final Benchmark correctly classified almost all RD readers, but misclassified more than half of proficient readers, with sensitivity and specificity of 96.4% and 43.9%, respectively. LG on ORF resulted in poor sensitivity and adequate specificity, at 53.6% and 93.6%, respectively. Classification under the DD criterion was identical to that under LG, since every child with LG also scored below the Final Benchmark cut-off for RD.

The 25th percentile cut-off on ORF yielded sensitivity of 71.4% and specificity of 83.5%. Under this criterion, 8 RD readers and 23 proficient readers were misclassified. The 33rd percentile criterion correctly identified an additional 2 RD readers and misclassified an additional 10 proficient readers, resulting in sensitivity and specificity of 78.6% and 76.3% respectively.

Chapter 5: Discussion

In this study, I examined the ability of 1st and 2nd grade reading measures to predict RD in 3rd grade for students who received *access* as needed to Tier II reading intervention from 1st to 3rd grade. I paired responsiveness criteria (e.g. Final Benchmark, Dual Discrepancy, 25th percentile) with predictive 1st and 2nd grade measures to identify measure/criteria combinations most effective in identifying children classified as RD by the beginning of 3rd grade. Finally, I explored the extent to which results replicated with a new cohort of students (i.e. Cohort B) whose instructional access (including schools,

teachers, and interventions) was nearly identical to the initial cohort (i.e. Cohort A), but who attended 1st grade the year after Cohort A. Participant data were drawn from a larger RtI study and were limited to those children who were present and had access to intervention from 1st to 3rd grade in the years 2007-2011.

Before exploring results of this study, it is important to address a few caveats that may explain differences between findings reported here and those of previous studies. First, this study differs from previous similar studies in that the participants for this study received access to Tier II intervention as needed from 1st to 3rd grade. Students' early reading scores collected in previous studies were in response to either general education (Tier I; Compton et al., 2006; Schatschneider, Wagner, & Crawford, 2008; Speece & Case, 2001) or intervention (Tier II; Compton et al., submitted; Simmons et al., 2008, Speece et al., 2003), and only those participants who were assigned to Tier I or Tier II instruction initially comprised the final samples. For example, participants in Simmons et al. were recruited for Tier II intervention in kindergarten; the initial kindergarten sample comprised the longitudinal sample that was followed through 3rd grade, and students who did not qualify for Tier II in kindergarten but may have qualified in later grades were excluded from analyses. Therefore, it is impossible to determine whether results of existing studies also pertain to students with late-emerging RD.

Also, I found no study that investigated the ability of early reading measures to predict RD based on distal outcomes when predictors were collected in response to the *most appropriate* instruction available within an RtI framework. Results yielded from samples inclusive of students without access to intervention may be obscured. To

illustrate this point, consider findings of Schatschneider et al. (2008). For a sample of students receiving Tier I instruction, growth in ORF during 1st grade explained no unique variance in 2nd grade reading fluency or comprehension skill after the end-of-1st grade ORF score were considered. If some participants required more intensive instruction than was offered in Tier I, Schatschneider et al.'s conclusion that growth is not a useful predictor when measures of final status are available, may be flawed. When students are not receiving appropriate instruction, their growth may be low and their final scores on RtI measures may also be low and therefore include information about growth, as Schatschneider et al. suggested. However, if the same students receive appropriate instructional access (i.e. access to intervention as needed), they may evidence good growth. Indeed, Fuchs et al. (2008) found steeper 1st grade WIF slopes for students who received intervention compared to controls, who did not receive the intensive instruction they needed to grow. So, final scores on a given measure obtained after access to appropriate instruction may or may not reflect growth, in which case measures of final status *and* growth may be needed to accurately measure students' responsiveness to instruction. In the case where initial scores were very low and growth was good, final scores may still be low despite good growth. Conversely, if initial scores were just below average and growth was good, final scores may concur with growth scores in classifying the child as a good responder to instruction. In all, by allowing the study sample to comprise students who were presumably receiving access to the most appropriate instruction available, early predictors and criteria may be more reliable in their designation of good and poor responder groups and in their prediction of later RD.

Additional ways in which this study differs from previous studies are that students were older when classified into RD and proficient reader groups (i.e. 3rd grade versus 1st or 2nd grade) and that classification was made according to scores on two latent constructs: Word Reading/Fluency and Comprehension/Vocabulary. Classifying older students into RD groups increases the likelihood that RD groups will contain late-emerging poor readers. And since approximately 88% of late-emerging poor readers may have deficits in either word reading or reading comprehension alone (Catts et al., 2011), classifying students with deficits in WR-F and C-V separately was necessary to determine whether early reading measures had power to identify early and late-emerging poor readers, whose deficits may occur only in specific constructs. These procedures for RD identification differ from studies where outcomes were collected in 1st or 2nd grade and were based on comprehensive reading constructs (i.e. a composite of word reading, fluency, and comprehension skill; e.g. Compton et al., submitted; Compton et al., 2010; Fuchs et al., 2008), but are similar to studies investigating predictors of late-emerging RD (e.g. Catts et al., 2011; Compton et al., 2008). In sum, because participants in this study were not limited to those who demonstrated risk in 1st grade (i.e. participants continued to contribute data regardless of whether they qualified for intervention in 1st grade, 2nd grade, 3rd grade, or not at all), analyses reflect the ability of early reading measures to predict RD classification in 3rd grade for students who showed early reading deficits in either WR-F or C-V, those who showed late-emerging deficits, and those who never showed deficits.

A final decision that may impact results involves the cut-off used to create RD and proficient reader groups. The cut-off used to identify RD readers in this study (i.e. 1.0 standard deviation below a locally normed mean) seemed reasonable since sample-referenced cut-points are commonly used in studies where groups of RD and proficient readers are dichotomized based on some outcome or group of outcomes (e.g. Catts et al., 2011; Speece, Schatscneider, Silverman, Case, Cooper, & Jacobs, 2011). Additionally, sample prevalence rates for each of the WR-F (sample prevalence = 13% Cohort A; 14% Cohort B) and C-V (sample prevalence = 14% Cohort A; 15% Cohort B) constructs in this study mirrored those of other studies using a 1.0 standard deviation cut-off to define RD (e.g. sample prevalence of RD was 15% in Compton et al., submitted). Furthermore, an investigation of differences on individual 3rd grade outcomes between RD and proficient reader groups under the 1.0 standard deviation cut-off provided support for this method, at least for the purpose of this study. RD and proficient readers had significantly different scores on individual outcomes for each of the WR-F and C-V constructs under the 1.0 standard deviation cut-off, and effect sizes for differences on construct-relevant assessments were generally larger than those on construct-secondary assessments. However, I recognize that the choice of a 1.0 standard deviation cut-off is arbitrary and may not reflect how students are identified as RD in other studies or in real-life contexts. Therefore, I explored group compositions across an additional plausible cut-off criterion.

Varying Criteria for RD Classification

To explore how results might differ given alternative cut-offs to define children as RD, I analyzed RD and proficient group compositions when defined by a cut-off of 1.5

standard deviations below the locally normed mean. Similar to the 1.0 cut-off, differences in individual 3rd grade outcomes between RD and proficient readers under the 1.5 cut-off were large and significant. Effect sizes for differences between groups on construct-relevant measures were larger than those for construct-secondary measures.

An exploration of means on individual 3rd grade outcomes by cut-off (1.0 vs. 1.5) suggested that the 1.5 cut-off identified a more severely impaired RD group, and logistic regression equation differences that arose from using the distinct cut-offs supported this notion. For the WR-F construct, 1st grade WIF was a significant predictor of RD under each cut-off, and the magnitude of the 1st grade WIF coefficient was larger under the 1.5 cut-off method. The same trend was found for 2nd grade spring ORF; that is, ORF was significant in each regression, with a larger magnitude in the regression using the 1.5 cut-off. For the C-V construct, 1st grade TOLD was the only significant 1st grade measure under each cut-off, and the magnitude of the coefficient was larger under the 1.5 cut-off. As expected, the RD group had the lowest means on predictors, and slope coefficients suggested that odds of being in the RD group were lower under the 1.5 cut-off, where very low scores were required for RD designation.

Another difference between the 1.0 and 1.5 cut-off regressions was the impact of 2nd grade predictors of RD in C-V. Although 2nd grade TOLD and ORF were significant predictors of RD under the 1.0 cut-off, they were not significant under the 1.5 cut-off when accounting for 1st grade TOLD. When 1st grade TOLD was removed from the analysis, however, coefficients on 2nd grade ORF and 2nd grade TOLD were significant. It may be that children with the most severe RD in 3rd grade also score the lowest among

their peers on vocabulary knowledge in 1st grade; severe early vocabulary deficits may prevent timely acquisition of early reading and comprehension skill, resulting in persistently low scores on these measures, which are also explained by 1st grade vocabulary deficits.

Another interesting difference in logistic regression equations between cut-offs was that the gender coefficient for RD classification was significant under the 1.0 cut-off, but not the 1.5 cut-off. The latter non-significant effect was likely due to low power, since RD sample sizes were small only for the 1.5 cut-off; the trend of overrepresentation of females in the RD group was consistent for 1.0 and 1.5 cut-offs. However interesting or informative the similarities and differences are in outcome means and logistic regression results for distinct RD cut-offs, the use of a 1.0 standard deviation cut-off served as a descriptive tool for this study, where the main purpose was to identify early reading measures and criteria for the identification of RD and not to identify the best criteria for defining RD in 3rd grade.

Significant 1st and 2nd Grade Predictors of RD in 3rd Grade

WIF. Some significant predictors of RD found in this study have also been reported in previous studies. Namely, 1st grade WIF has shown promise as a screening tool (Compton et al., 2006; 2010; Fuchs et al., 2008; Speece et al., 2011) and gauge of intervention responsiveness (Fuchs et al., 2003) in past research, and was also a significant predictor of WR-F RD in this study. Also, similar to Compton et al. (2010), 1st grade ORF was not useful in the prediction of WR-F reading skill after accounting for 1st grade WIF. Floor effects typically found on ORF at the 1st grade level may account

for its poor predictive capability (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009). At least when predicting distal deficits in WR-F using 1st grade measures, WIF appears to outperform ORF.

Although 1st grade WIF was a significant predictor of 3rd grade RD in the logistic regression model, its usefulness as an isolated predictor was limited. No responsiveness criteria paired with WIF provided optimal sensitivity for predicting RD. Under criteria of .90 for sensitivity and .80 for specificity, the 33rd percentile criterion performed best, correctly classifying 82% of RD and 76.1% of proficient readers in Cohort A with comparable sensitivity and specificity for Cohort B (sensitivity= 88.5%, specificity= 76.1%). Although the 33rd percentile criterion used in this study identified most readers who would become proficient or RD readers by 3rd grade, some readers from each group were misclassified.

Differences in the usefulness of WIF for predicting RD between this and other studies may stem from differences in criteria applied to WIF. Growth estimates for WIF were not explored here, but have been found useful for identifying children with persistent reading deficits at the end 1st (Speece et al., 2011) and 2nd (Compton et al., 2006) grade; future research might explore effects of WIF growth data on RD classification made past grade 2. In all, even though the 1st grade WIF/percentile measure/criterion combination may not serve optimally as a sole predictor of RD in 3rd grade, employing WIF as a screening or progress monitoring measure in 1st grade may help identify children who need more intensive intervention and may weed out those who are performing well given current instruction.

Woodcock Reading Mastery Tests: Passage Comprehension. Interestingly, 2nd WRMT PC raw scores were also significant predictors of 3rd grade RD group membership within the WR-F construct, even after accounting for effects of 1st grade WIF and 2nd grade ORF. Perfetti's (1985) Lexical Quality Hypothesis (LQH) provides a viable explanation as to why this may be. The LQH states that proficient reading requires high quality phonemic, orthographic, and semantic representations of words. Research on the LQH suggests that good comprehenders typically have higher quality phonemic and orthographic word representations compared to poor comprehenders (Perfetti & Hart, 2003). There is ample evidence to support a directional relationship from reading fluency skill to subsequent comprehension skill, and empirical evidence for the reverse is growing. For example, Jenkins et al. (2003) found that for 4th graders with a range of reading ability, reading comprehension skill uniquely predicted context fluency, even after controlling for single word reading fluency. In a more recent study where researchers used path analysis to examine measurement and structural relations among early reading skills for 2nd grade students, Hudson et al. (2012) found evidence for a direct effect of reading comprehension on text reading fluency, even after accounting for direct effects of decoding fluency and single word reading fluency. Similar to the current study, the coefficient from reading comprehension to reading fluency in Hudson et al. was small (b= .18) compared to word reading fluency skills [decoding fluency (b=.20) and single word reading fluency (b=.74)], but the effect was significant. In all, evidence from the current study suggests that in addition to predicting variance in reading fluency

outcomes, passage comprehension might also play a small role in the prediction of WR-F RD for 3rd grade students.

Although 2nd grade PC was a significant model-based predictor of RD, it fared less well as a stand-alone predictor. The 33rd percentile criterion provided the best balance between sensitivity and specificity for each Cohort, with sensitivity and specificity for Cohort A of 79.3% and 81.6%, respectively, and for Cohort B of 80% and 83.6%, respectively. Second grade PC better predicted proficient 3rd grade readers, which seems useful for screening purposes (i.e. to weed out students without need of Tier II); however, passage comprehension measures are time consuming to administer compared to other measures (e.g. ORF) that provide equivalent specificity when paired with useful criteria. Nevertheless when available, passage comprehension may add power to models specified to predict RD in WR-F. Employing passage comprehension measures in "gated" screening procedures (Compton et al., 2010) may be an example of such use.

ORF. The last measure to which criteria were applied in forming groups of good and poor responders was ORF. Not surprisingly, ORF collected at the end of 2nd grade was a significant predictor of RD in WR-F -- research consistently shows that RD readers lag behind their proficient peers on measures of reading fluency. When paired with RtI criteria, however, the ability of ORF to predict RD varied. Applying the 2nd grade Final Benchmark criterion (90 wcpm) to ORF scores resulted in excellent sensitivity and poor specificity for Cohorts A and B. As found in other intervention studies (O'Connor & Jenkins, 1999; Fuchs et al., 2004), published benchmarks appeared overly stringent and misclassified typically developing readers as poor responders to instruction. When the

Final Benchmark score was reduced to the highest score needed to capture all RD readers, specificity improved. For Cohort A, a Final Benchmark score of 75 resulted in sensitivity of 100% and specificity of 78.8%. For Cohort B, a Final Benchmark of 70 wcpm resulted in sensitivity of 100% and improved, yet still poor, specificity (Specificity= 69.1%). Overall, a Final Benchmark of 90 wcpm on ORF at spring of 2nd grade falls short as a stand-alone predictor of RD. Sample-derived Final Benchmarks (Cohort A= 75; Cohort B = 70) yielded improved specificity and excellent sensitivity, and fall within the "some risk" category according to DIBELS. These benchmarks may be more useful in RD identification attempts.

In contrast to findings reported in Fuchs et al. (2004), Low Growth and DD criteria performed poorly when predicting RD in WR-F. For the current study, correct classification of RD students never exceeded 70% when ORF was paired with Low Growth or DD criteria. As previously mentioned, many students receiving access to intervention as needed might be expected to yield adequate growth on measures in response to instruction. However, some children who were good responders under Low Growth and/or DD on ORF were designated RD by 3rd grade, resulting in low sensitivity rates in the current study. Burns and Senesac (2005) reported a similar phenomenon when using ORF scores to predict outcomes on a standardized reading assessment. The researchers found that slope estimates did little to differentiate students on reading outcomes when slopes were examined within a subsample of 2nd graders whose final ORF scores fell below the final benchmark cut-point. So even though some students responded well (i.e showed adequate growth) to Tier II instruction and others did not (i.e.

they showed poor growth), students' scores on outcomes were indistinguishable. This phenomenon may have also been operating in this study, and may explain why the ORF/DD measure/criterion combination was unable to accurately classify students into RD and proficient reader groups. When children with reading deficits demonstrate reading growth commensurate with proficient readers, they are indeed responding to instruction. However, commensurate growth may not be enough to shield very poor readers from persistent reading problems, especially as reading demands change in 3rd and 4th grade. Therefore, researchers and practitioners must be careful when using Low Growth or DD criteria to disqualify students for more intensive intervention – adequate growth in response to current instruction does not necessarily imply that growth is great enough to result in future proficient reading, especially past 2nd grade. The gap between poor readers' skills and those of their typically developing peers might remain.

Another explanation of why DD fared less well as a criterion for determining good or poor response status pertains to differences in how reading outcomes were operationalized in this study and in others. Outcomes in studies supporting the use of DD to describe responsiveness were continuous scores on commonly used reading assessments (Fuchs et al., 2004; Speece & Case, 2001). In these studies, DD students' outcome scores were significantly lower than scores for students who had good growth, but did not meet final achievement benchmarks. In the current study, students were classified as RD if their composite scores on WR-F (comprising decoding, word identification, reading fluency, and spelling tasks) fell 1.0 standard deviation below the sample mean. If DD classifies the *most* impaired readers on some skill, then the low

sensitivity rates found in this study may be explained in a few ways. First, many students in the current study who were *not* DD still scored 1.0 standard deviation below the sample mean on WR-F, perhaps because the 1.0 standard deviation cut-off did not stringently identify only the most impaired readers. Also, some students who *were* DD had scores on other measures of reading that met or exceeded those of their non-DD peers. Each of these scenarios may have led to poor classification accuracy. So although a DD criterion typically identifies the poorest readers on the measure to which it is applied, it may not identify *all* readers who will eventually develop RD, may over-select for students who struggle mostly on a particular skill, and may miss those students with deficits in other areas of reading. Additionally, because DD may only identify students who improve the least during the classification years (i.e. 1st and 2nd grade when students receive Tier II intervention), the DD criterion may misidentify late-emerging poor readers if used as the sole criterion to identify poor response to instruction. Readers with late-emerging RD tend to have lower scores than their peers in early grades, but their scores are not distinct enough to cause alarm or be used as accurate predictors for later RD (Catts et al., 2011; Compton et al., 2008). Therefore, although the ORF/DD measure/criterion combination might be a useful for identifying the most impaired readers on ORF, many of whom will eventually develop RD, the criterion may be too selective and may overlook students with less severe RD and those who develop late-emerging RD.

In addition to being a significant predictor of RD in WR-F, 2nd grade ORF was also a significant predictor of RD in the logistic model for C-V. As noted, the relation

between reading fluency and comprehension is well established as has been demonstrated in many studies of reading. One theory that may explain the impact of fluent reading on comprehension is Perfetti's (1985) Verbal Efficiency Theory. Perfetti noticed that groups comprised of good and poor comprehenders also tended to be naturally separated by good and poor lexical skill, and posited that many problems in reading comprehension stem from poor word reading skill (Perfetti & Hart, 2003). Clearly, students will have great difficulty understanding what they read if the act of word reading is also strained. Recent research with a sample of 4th grade students with a range of reading abilities showed that reading fluency skill was a strong predictor of reading comprehension, even after accounting for word reading skill (Fuchs et al., 2003). So, the significant effect of ORF on RD in C-V was predictable and provides further support for supplemental instruction in reading fluency as needed to improve reading comprehension, at least for students in 2nd and 3rd grade.

Although model sensitivity and specificity for predicting RD in C-V approached acceptable rates when 1st and 2nd grade predictors were entered in logistic regression analyses, no stand-alone predictor adequately predicted RD in C-V when paired with any criteria explored. For example, the best sensitivity rates for ORF were found under the Final Benchmark criterion (Cohort A sensitivity = 81.3%, Cohort B sensitivity= 96.4%). However, classification of proficient readers was unacceptable (Cohort A specificity = 54.2%; Cohort B specificity = 43.9%). In contrast to results under the WR-F model, ORF also fared poorly when paired with percentile-cut criteria. The highest sensitivity rate across cohorts was 78.6%, with specificity of 76.3%. First and second grade TOLD

paired with percentile criteria also produced low sensitivity and specificity rates for each cohort in general, with no criteria providing sensitivity above .8. Nevertheless a model comprising gender, 2nd grade ORF, and 2nd grade TOLD correctly classified over 80% of RD readers in each Cohort, and between 70% and 76% of Proficient readers in Cohorts B and A, respectively.

Overall, prediction of RD in C-V was less successful than prediction of RD in WR-F. Early reading scores for late-emerging RD readers may have influenced prediction accuracy for RD in C-V especially. Specifically, late-emerging poor readers may have been classified as good responders to instruction in 1st and 2nd grade, but classified as RD in comprehension in 3rd grade, causing sensitivity rates to plummet. It is possible that although students with and without RD in comprehension differed on early reading measures, differences in their scores on these measures were not powerful enough to classify students into RD and proficient reader groups. Evidence from Catts et al. (2011) supports this notion. Like the current study, Catts et al. experienced poor ability to accurately classify students with RD in comprehension using early reading measures. Although students with and without RD in comprehension differed on kindergarten measures (including vocabulary and narrative comprehension and recall), no early measures were sensitive enough to correctly classify more than 80% of readers who would eventually develop RD in comprehension. In fact, prediction accuracy was as low as 16% for readers with late-emerging RD in comprehension and for combined RD in word reading and comprehension. On a positive note, the model-based approach to classifying students with RD in comprehension provided improved sensitivity and

specificity rates for students in this sample. When using a model comprising gender, 2nd grade TOLD, and 2nd grade ORF, sensitivity and specificity rates were approximately 82% and 70% for students in each Cohort. Therefore, model-based approaches may be the best way to identify children in 1st and 2nd grade who will show deficits in C-V in 3rd grade.

Models vs. Single Measures

Support of model-based approaches (as opposed to single-measure approaches) for identifying children as good or poor responders to instruction is strong in the literature (Compton et al., 2010; O'Connor & Jenkins, 1999). The current study lends support to the notion that model-based approaches are typically more powerful and reliable predictors of distal RD, especially for RD in C-V. Indeed, sensitivity and specificity rates for correct classification were consistently higher within logistic regression models compared to most single-measure predictors. It is interesting to note, however, that the 33rd percentile criterion on 1st grade WIF and the 25th percentile criterion on 2nd grade ORF produced sensitivity and specificity rates that approached those of the model-based rates. WIF was a less accurate predictor of distal RD than ORF, perhaps because comparatively more time elapsed between the WIF administration and 3rd grade RD classification. Other researchers have noted that reading measures collected temporally nearer outcomes tend to be stronger predictors of reading outcomes than similar measures collected earlier (O'Connor & Jenkins, 1999). For ORF, sensitivity and specificity rates fell short of those produced by the model-based rates for predicting RD in WR-F by just 1% and 5%, respectively within Cohort A. Within Cohort B, sensitivity rates were higher

by 4.6% and specificity fell just below that of the model-based estimate. Overall, ORF scores collected in spring of 2nd grade may be a viable substitute for a model-based approach in the classification of good and poor responder groups and/or prediction of WR-F skill in 3rd grade. This approach is promising given the time consuming nature of collecting multiple measures, but is limited in that prediction is only powerful within the WR-F construct.

Replication

A persistent need in intervention research involves replication of models for RD classification on a new sample of students. Models are typically developed and tested on a single sample, which may produce biased classification results and requires replication on a new sample to generate confidence in study findings. The replication of regression models and measure/criteria combinations used for RD classification in the current study was impressive. It is important to note that students in each of the cohorts received instruction in similar environments (same schools, teachers, access to Tier II, and Tier II content when applicable), but differed in that students in Cohort B entered 1st grade the year following those in Cohort A. Therefore, replication was conducted on a sample very similar to the descriptive sample. Indeed, differences between cohorts in sensitivity and specificity indices for logistic regressions were small, with differences ranging from less than 1% to 7%. Differences in sensitivity and specificity indices across Cohorts A and B were also small for measure/criteria combinations deemed useful in the prediction of RD, and fell between 1% and 5%. That the models developed with the original sample of

students performed equally well when applied to a second sample provides support for many conclusions drawn in this study.

Limitations

Some features of this study that act as extensions to the literature may also serve as limitations. First, participants in this study received *access* to a high-quality Tier II intervention from 1st to 3rd grade. Some participants qualified for and received such intervention while others did not qualify for, and therefore did not receive, intervention. Yet, all students' scores were used in model-based and single-measure based predictions. This feature of the current study extends previous research in that early predictors of distal RD were collected under the assumption that students were receiving access to the most appropriate instruction available – the case where a 1st grade poor reader has low scores due to inappropriate access to instruction is less likely. Although poor performance from participants in this study may still be a function of the instructional environment (i.e. even Tier II is not intensive enough to meet their needs), it is equally likely that poor responders in this study were truly at risk for later RD and would continue to be at risk since they were already participating in a functional RtI framework. That being said, results of this study hinge on the availability of intervention for struggling students in 1st and 2nd grade. Model-based and singular predictors of RD that showed promise in this study may be less powerful or inappropriate to use in studies or settings where children do not have fluid access to secondary intervention.

It is also important to note that participants in this study attended high poverty schools and many spoke English as a second language; however, instruction provided in

Tier I and Tier II may have been of higher quality than that offered in other similar settings. General education teachers in this study received 120 hours of professional development in the delivery of best-practice reading instruction. Also, many Tier II tutors were graduate students or former special education teachers, and most had experience working with children at-risk for reading delays. Thus, the Tier I and II instructional environments in this study may not reflect those in other low-income, high-minority schools. The extent to which results gained from this study apply to typical low-income instructional settings should be explored.

Practical Implications and Future Directions

Response-to-intervention has enjoyed support as an early intervention framework, but sparse research supports its role as an LD identification tool. Results from this study and others show that groups of good and poor responders vary depending on measures and criteria used to classify children into groups. Furthermore, prediction accuracy of distal RD based on early intervention measures hinges on the choice of predictors and on the definition of RD. Even with these considerations, some early reading measures show promise as indicators of future RD, at least in the areas of word reading and reading fluency.

First, word identification fluency measures collected at the end of 1st grade may help teachers identify students in need of additional reading support, regardless of the level of support currently provided. Additionally, text reading fluency measures may help 2nd grade teachers isolate those students who will need additional reading support before the start of 3rd grade. In settings where intervention is continuously offered, these

early reading assessments show promise for identifying readers who will need more or different reading instruction to improve.

Conversely, additional effort should be focused on identification of students who will show persistent difficulty in reading comprehension. Currently, early reading measures do little to differentiate between students who will and will not become proficient comprehenders of text. According to results from this study, multivariate methods may be needed to adequately identify children who will struggle with comprehension; thus, additional teacher training and time may be required to effectively translate such approaches into classroom practice. Clearly, further research is required to determine whether children with late-emerging RD can be identified using early screening measures. Failing to identify late-emerging poor readers before severe deficits develop will perpetuate the "wait-to-fail" cycle for these students, specifically.

References

Allison, P.D. (2002). *Missing data.* Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.

Al Otaiba, S. & Fuchs, D. (2006). Who are the young children for whom best practices are ineffective? An experimental and longitudinal study. *Journal of Learning Disabilities, 39*(5), 414-431.

Barth, A.E., Stuebing, K.K., Anthony, J.L., Denton, C.A., Mathes, P.G., Fletcher, J.M. & Francis, D.J. (2008). Agreement among response to intervention criteria for identifying responder status. *Learning and Individual Differences, 18,* 296-307.

Bradley, R., Danielson, L., & Hallahan, D. (Eds.). (2002). *Identification of learning disabilities: Research to practice*. Mahwah NJ: Erlbaum

Brown Waesche, J.S., Schatschneider, C., Maner, J.K., Ahmed, Y., & Wagner, R. (2011). Examining agreement and longitudinal stability among traditional and RTI-based definitions of reading disability using the affected-status agreement statistic. *Journal of Learning Disabilities, 44*(3), 296-307.

Burns, M.K., & Sensac, B.V. (2005). Comparison of dual discrepancy criteria to assess response to intervention. *Journal of School Psychology, 43,* 393-406.

Cain, K., & Oakhill, J. (2011). Matthew effects in young readers: Reading comprehension and reading experience aid vocabulary development. *Journal of Learning Disabilities, 44*(5), 431-443.

Catts, H.W., Compton, D., Tomblin, J.B., & Sittner Bridges, M. (2011). Prevalence and nature of late-emerging poor readers. *Journal of Educational Psychology, 104*(1), 166-181.

Catts, H.W., Petscher, Y., Schatschneider, C., Bridges, M., & 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.

Chall, J.S. (1996). *Stages of reading development* (2nd ed.). Fort Worth, TX: Harcourt Brace.

Chall, J.S., Jacobs, V.A., & Baldwin, L.E. (1990). *The reading crisis: Why poor children fall behind.* Cambridge, MA: Harvard University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/ correlational analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Erlbaum.

Compton, D.L., Fuchs, D., Fuchs, L.S., Bouthon, B., Gilbert, J., Barquero, L.A., … Crouch, R.C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology, 102*(2), 327-340.

Compton, D.L., Fuchs, D., Fuchs, L.S, Bryant, J. (2006). Selecting at-risk reading in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98*(2), 394-409.

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

Compton, D.L., Gilbert, J.K., Jenkins, J.R., Fuchs, D., Fuchs, L.S, Cho, E., … Bouton, B.D. (submitted). Fast-tracking chronically unresponsive children into tier 3 instruction: What level of data is necessary to ensure adequate selection accuracy?

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

Donovan, M. S. & Cross, C. T. (2002). *Minority Students in Special and Gifted Education*. Washington, DC: National Academy Press.

Dunn, L. M., Dunn, L. M., & Dunn, D. M. (1997). *The Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Services, Inc.

Ehri, L.C. (1998). Grapheme–phoneme knowledge is essential for learning to read words in English. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 3–40). Mahwah, NJ: Erlbaum.

Ehri, L.C. (2005). Learning to read words: Theory, findings and issues. *Scientific Studies of Reading, 9*(2), 167-188.

Elbaum, B., Vaughn, S., Hughes, M., & Moody, S.W. (1999). Grouping practices and reading outcomes for students with disabilities. *Exceptional Children, 65*(3), 399-415.

Fletcher, J.M., Shaywitz, S.E., Shankweiler, D.P., Katz, L., Liberman, I.Y., Stuebing, K.K…Shaywitz, B.A. (1994). Cognitive profiles of reading disability: Comparisons of discrepancy and low achievement definitions. *Journal of Educational Psychology, 86*(1), 6-23.

Foorman, B., Schatschneider, C., Eakin, M.N., Fletcher, J.M., Moats, L.C., & Francis, D.R. (2006). The impact of instructional practices in grades 1 and 2 on reading and spelling achievement in high poverty schools. *Contemporary Educational Psychology, 31,* 1-29.

Francis, D.J., Fletcher, J.M., Stuebing, K.K., Reid Lyon, G., Shaywitz, B.A., & Shaywitz, S.E. (2005). Psychometric approaches to the identification of LD: IQ and achievement scores are not sufficient. *Journal of Learning Disabilities, 38*(2), 98-108.

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

Fuchs, D. & Fuchs, L.S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly, 41*(1).

Fuchs, L. S. & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research & Practice, 13,* 204– 219.

Fuchs, D., Compton, D.L, Fuchs, L.S., Bryant, J., & Davis, N.G. (2008). Making "secondary intervention" work in a three-tier responsiveness-to-intervention model: findings from the first-grade longitudinal reading study of the National Research Center on Learning Disabilities. *Reading and Writing, 21,* 413-436.

Fuchs, D., Fuchs, L.S., & Compton, D.L. (2004). Identifying reading disabilities by responsiveness-to-intervention: Specifying measures and criteria. *Learning Disability Quarterly, 27*(4), 216-227.

Fuchs, D., Mock, D., Morgan, P.L., & Young, C.L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research & Practice, 18*(3), 157-171.

Gersten, R., Baker, S.K., Haager, D., & Graves, A.W. (2005). Exploring the role of teacher quality in predicting reading outcomes for first-grade English learners: An observational study. *Remedial and Special Education, 26(4),* 197-206.

Gresham, F. (2002). Reponsiveness to intervention: An alternative approach to the identification of learning disabilities. In Bradley, R., Danielson, L., & Hallahan, D. (Eds). *Identification of learning disabilities: Research to practice.* Mahwah NJ: Erlbaum

Good, R.H., & Kaminski, R.A. (2003). *Dynamic Indicators of Basic Early Literacy Skills 6th Edition* (6th ed.). Longmont, CO: Sopris West Educational Services.

Good, R. H., Kaminski, R. A., Shinn, M., Bratten, J., Shinn, M., Laimon, L., Smith, S., & Flindt, N. (2004). *Technical Adequacy and Decision Making Utility of DIBELS (Technical Report No. 7*). Eugene, OR: University of Oregon.

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

Hammill, D.D. & Newcomer, P.L. (2008). *Test of Language Development* (4th ed.). Austin, TX: Pro-Ed, Inc.

Harry, B. & Klingner**,** J. K. (2006). *Why are so many minority students in special education? Understanding race and disability in schools.* New York : Teachers College Press.

Hart, B. & Risley, T.R. (1995). *Meaningful differences in the everyday experience of young American children.* Baltimore, MD: Paul H Brookes Publishing.

Hudson, R.F., Torgesen, J.K., Lane, H.B., & Turner, S.J. (2012). Relations among reading skills and sub-skills and text-level reading proficiency in developing readers. *Reading and Writing, 25*(2), 483-507.

Ikeda, M.J., Rahn-Blaskeslee, A., Neibling, B.C., Gustafson, J.K., Allison, R., & Stumme, J. (2007). The Heartland Area Education Agency 11 Problem-Solving Approach: An Overview and Lessons Learned. In S.R. Jimerson, M.K. Burns, & VanDerHeyden, A.M. (Eds). *Handbook of Response to Intervention: The science and practice of assessment and intervention* (pp. 225-278). New York, NY: Springer Science+Business Media.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology, 95*(4), 719-729.

Juel, C. & Minden-Cupp, C. (2000). Learning to read words: Linguistic units and instructional strategies. *Reading Research Quarterly, 35*(4), 458-492.

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

Kamps, D., Abbott, M., Greenwood, C., Wills, H., Veerkamp, M., & Kaufman, J. (2008). Effects of small-group reading instruction and curriculum differences for students most at risk in kindergarten: Two-year results for secondary- and tertiary-level interventions. *Journal of Learning Disabilities, 41*(2), 101-114.

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

Klingner, J.K., Vaughn, S., & Boardman, A. (2007) Teaching reading comprehension to students with learning difficulties. In Karen R. Harris & Steve Graham (Eds). *What works for special-needs learners.* New York, NY: The Guilford Press.

Larsen. S., Hammill, D.D., & Moats, L. (1999). *Test of Written Spelling: Examiner's Manual*. Austin, TX: Pro-Ed, Inc.

Lenz, B.K. & Hughes, C.A. (1990). A word identification strategy for adolescents with learning disabilities. *Journal of Learning Disabilities, 23*(3), 149-163.

Lerner, J. (1989). Educational intervention in learning disabilities. *Journal of the American Academy of Child and Adolescent Psychiatry,* 28, 326-331.

Lerner, J. (2003). *Learning disabilities: Theories, diagnosis, and teaching strategies* (9th ed.). Boston: Houghton-Mifflin.

Mathes, P. G., Denton, C.A., Fletcher, J.M., Anthony, J.L., Francis, D.J., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics on the skills of struggling readers. *Reading Research Quarterly, 40*(2), 148-182. doi: 10.1598/RRQ.40.2.2

Mathes, P. G., Torgesen, J. K., & Allor, J. H. (2001). The effects of peer-assisted literacy strategies for first-grade readers with and without additional computer-assisted instruction in phonological awareness. *American Educational Research Journal, 38*(2), 371-410. doi: 10.3102/00028312038002371

McMaster, K.L., Fuchs, D., Fuchs, L.S., & Compton, D.L. (2005). Responding to nonresponders: An experimental field trial of identification and intervention methods. *Exceptional Children, 71*(4), 445-463.

Mellard, D.F., McKnight, M., & Woods, K. (2009). Response to intervention screening and progress-monitoring practices in 41 local schools. *Learning Disabilities Research & Practice, 24*(4), 186-195.

Morris, D., Trathen,W., Lomax, R.G., Perney, J., Kucan, L., Frye, E. M., …Schlagal, R. (2012). Modeling aspects of print-processing skill: implications for reading assessment. *Reading and Writing, 25*(1), 189-215.

National Reading Panel. (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.* Washington, DC: National Institute of Child Health and Human Development.

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

Newcomer, P.L. & Hammill, D.D. (1997). *Test of Language Development* (3rd ed.). Austin, TX: Pro-Ed, Inc.

O'Connor, R. E. (2000). Increasing the intensity of intervention in kindergarten and first grade. *Learning Disabilities Research and Practice, 15*(1), 43-54.

O'Connor, R.E. (2007). Teaching word recognition. In Karen R. Harris & Steve Graham (Eds). *What works for special-needs learners*. New York, NY: The Guilford Press.

O'Connor, R.E., Bell, K.M., Harty, K.R., Larkin, L.K., Sackor, S., & Zigmond, N. (2002). Teaching Reading to Poor Readers in the Intermediate Grades: A Comparison of Text Difficulty. *Journal of Educational Psychology, 94*, 474-485.

O'Connor, R. E., Fulmer, D., Harty, K. R., & Bell, K. M. (2005). Layers of reading intervention in kindergarten through third grade: Changes in teaching and student outcomes. *Journal of Learning Disabilities, 38*(5), 440-455. doi: 10.1177/00222194050380050701

O'Connor, R. E. & Jenkins, J.R. (1999). Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading*, *3*(2), 159-197. doi: 10.1207/s1532799xssr0302_4

O'Connor, R.E., Sanchez, V., & Bocain, K. (submitted). Access to a responsiveness to intervention model: Does beginning intervention in kindergarten matter?

O'Connor, R.E., Swanson, L., & Geraghty, C. (2010). Improvement in reading rate under independent and difficult text levels: Influences on word and comprehension skills. *Journal of Educational Psychology*, *102*(1), 1-19.

O'Malley, K.J., Francis, D.J., Foorman, B.R., Fletcher, J.M., & Swank, P.R. (2002). Growth in precursor and reading-related skills: Do low-achieving and IQ-Discrepant readers develop differently*? Learning Disabilities Research and Practice, 17*(1), 19-34.

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

Perfetti, C.A. & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds), *Studies in Written Language and Literacy: Precursors of Functional Literacy,* 11, (pp. 189-213). Philadelphia, PA: John Benjamins Publishing Company.

Raudenbush, S.W. & Byrk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage Publications, Inc.

Raudenbush, S.W., Byrk, A.S., Cheong, Y.F, Congdon, R. T., & Toit, M. (2011). *HLM 7: Linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International, Inc.

Raykov, T. & Marcoulides, G. (2008). *An introduction to applied multivariate analysis.* New York, NY: Taylor and Francis Group.

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

Schnatschneider, C., Wagner, R., & Crawford, E. (2008). The importance of measuring growth in response to intervention models: Testing a core assumption. *Learning and Individual Differences, 18,* 308-315.

Schwanenflugel, P. J., Meisinger, E.B., Wisenbaker, J.M., Kuhn, M.R., Strauss, G.P., & Morris, R.D. (2006). Becoming a fluent and automatic reading in the early elementary school years. *Reading Research Quarterly*, *41*(4), 496-522.

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

Simmons, D.C., Hairrell, A., Edmonds, M., Vaughn, S., Larsen, R., Wilson, V. , … Byrns, G. (2010). A comparison of multiple-strategy methods: Effects on fourth-grade students' general and content-specific reading comprehension and vocabulary development. *Journal of Research on Educational Effectiveness, 3*(2), 121-156.

Speece, D.L. & Case, L.P. (2001). Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology, 93*(4), 735-749.

Speece, D.L., Case, L.P., & Molly, D.E. (2003). Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities Research &Practice, 18*(3), 147-156.

Speece, D.L., Schatscneider, C., Silverman, R., Case, L, 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.

Spicer, J. (2005). *Making sense of multivariate data analysis.* Thousand Oaks, CA: Sage Publications.

Stanovich, P. J. & Jordan, A. (1998). Canadian teachers' and principals' beliefs about inclusive education as predictors of effective teaching in heterogeneous classrooms. *The Elementary School Journal, 98*(3), 221-238.

Thurlow, M.L. Ysseldyke, J.E., Wotruba, J.W. & Algozzine, B. (1993). Instruction in special education classrooms under varying student-teacher ratios. *The Elementary School Journal, 93*(3), 305-320.

Torgesen, J.K., Alexander, A.W., Wagner, R.K., Rashotte, C.A., Voeller, K.K.S., & Conway, T. (2001). Intensive Remedial Instruction for Children with Severe Reading Disabilities: Immediate and Long-term Outcomes From Two Instructional Approaches. *Journal of Learning Disabilities, 34*(1), 33-58. doi: 10.1177/002221940103400104

Torgesen, J. K., Wagner, R. K., & Rashotte, C.A. (1997). Test of Word Reading Efficiency. Austin, TX: Pro-Ed.

Vadasy, P.F., Jenkins, J.R., Antil, L.R., Wayne, S.K., & O'Connor, R.E. (1997). The effectiveness of one-to-one tutoring by community tutors for at-risk beginning readers. *Learning Disability Quarterly, 20*, 126-139.

Vadasy, P. F., Sanders, E.A., & Tudor, S. (2007). Effectiveness of paraeducator-supplemented individual instruction: Beyond basic decoding skills. Journal of *Learning Disabilities, 40*, 508-525.

Vadasy, P.F., Wayne, S. K., O'Connor, R.E., Jenkins, J.R., Pool, K., Firebaugh, M., & Peyton, J. (2005). *Sound Partners: A tutoring program in phonics-based early reading.* Longmont, CO: Sopris West.

Vaughn, S., Linan-Thompson, S., & Hickman, P. (2003). Response to instruction as a means of identifying students with reading/learning disabilities. *Exceptional Children, 69*(4), 391-409.

Vaughn, S., Linan-Thompson, S., Kouzekanani, K., Bryant, D., Dickson, S. & Blozis, S. (2003). Reading instruction grouping for students with reading difficulties. *Remedial and Special Education, 24*(5), 301-315.

Vaughn, S., Linan-Thompson, S., Woodruff, A., Murray, C., Wanzek, J., Scammacca, N., Roberts, G., & Elbaum, B. (2008). Effects of professional development on improving at- risk students' performance in reading. In C.R. Greenwood, T.R. Kratochwill, & M. Clements (Eds), *Schoolwide prevention models: Lessons learned in elementary schools* (pp. 115- 141). New York, NY: The Guilford Press.

Vaughn, S., Wanzek, J., Murray, C.S., Scammaca, N., Linan-Thompson, S., & Woodruff, A.L. (2009). Response to early reading intervention: Examining high and low responders. *Exceptional Children, 75*(2), 165-183.

Vaughn, S., Wanzek, J., Woodruff, A., & Linan-Thompson, S. (2007). A three-tier model for preventing reading difficulties and early identification of students with reading disabilities. In D.H. Haager, S. Vaughn, & J.K. Klingner (Eds), *Validated reading practices for three tiers of intervention* (pp. 11-27). Baltimore: Brooks.

Vellutino, F.R., Scanlon, D.M., Sipay, E.R., Small, S.G., Chen, R., Pratt, A., & Denckla, M.B. (1996). Cognitive profiles of difficult to remediate and readily remediated poor readers: Early interventions as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88*(4), 601-638.

Vellutino, F.R., Scanlon, D.M., Zhang, H., & Schatschneider, C. (2008). Using response to kindergarten and first grade intervention to identify children at-risk for long-term reading difficulties. *Reading and Writing, 21*, 437-480.

Verhoeven, L., van Leeuwe, J., Vermeer, A. (2011). Vocabulary growth and reading development across the elementary school years. Scientific *Studies of Reading, 15*(1), 8-25.

Wanzek, J. & Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *School Psychology Review, 36*(4), 541-561.

Wanzek, J. & Vaughn, S. (2008). Response to varying amounts of time in reading intervention for students with low response to intervention. Journal *of Learning Disabilities, 41*(2), 126-142.

Weiser, B. & Mathes, P. (2011). Using encoding instruction to improve the reading and spelling performances of elementary students at risk for literacy difficulties: A best-evidence synthesis. *Review of Educational Research, 81*(2), 170-200.

Wiederholt, J.L., & Bryant, B.R. (2001). *Gray Oral Reading Tests* (4th ed.). Austin, TX: Pro-Ed.

Woodcock, R. (1998). *Woodcock Reading Mastery Test–Revised/Normative Update*. Circle Pines, MN: American Guidance Service.

Tables

Table 1
District Demographics

	District A		District B	
	n	%	*n*	%
Enrollment	56,727		19,987	
Ethnicity				
African American	9,224	16.3	888	4.4
Hispanic	38,605	68.1	14,352	71.8
White	6,164	10.9	3,073	15.4
American Indian	427	<1	78	<0.5
Asian	457	<1	666	3.3
Filipino	236	<0.5	331	1.7
Pacific Islander	186	<0.5	92	<0.5
Multiple/no response	1,741	3.1	507	2.5
Special Education	5,239	9.23	1,746	8.7
English Language Learners	24,848	43.8	10,068	50.4
Socioeconomically Disadvantaged	41,293	72.8	12,906	64.6

Table 2
School Demographics

	District A			District B	
	School A	School B	School C	School Y	School Z
Enrollment *(N)*	910	406	475	477	601
Ethnicity *n*(%)					
African American	165 (18.1%)	64 (15.8%)	84 (17.7%)	18 (3.8%)	25 (4.2%)
Hispanic	577 (63.4%)	279 (68.7%)	309 (65.1%)	379 (79.5%)	459 (76.4%)
White	110 (12.1%)	35 (8.6%)	48 (10.1%)	57 (11.9%)	78 (13.0%)
American Indian	5 (0.5%)	4 (1.0%)	3 (0.6%)	1 (0.2%)	1 (0.2%)
Asian	6 (0.6%)	0 (0%)	4 (0.8%)	5 (1.0%)	22 (3.7%)
Filipino	2 (0.2%)	1 (0.2%)	1 (0.2%)	0 (0%)	10 (1.7%)
Pacific Islander	11 (1.2%)	1 (0.2%)	7 (1.5%)	0 (0%)	3 (0.5%)
Multiple/no response	34 (3.7%)	22 (5.4%)	19 (4.0%)	17 (3.6%)	3 (0.5%)
Special Education	Not Available	Not Available	Not Available	Not Available	Not Available
English Language Learners	310 (34.5%) 95% Spanish	138 (33.9%) 99% Spanish	156 (32.8%) 98% Spanish	285 (59.7%) 97% Spanish	314 (52.2%) 94% Spanish

Table 3
Descriptive Statistics by Cohort

	Cohort A (0708 1st) *n* (proportion)	Cohort B (2008-2009 1st) *n* (proportion)	
Total N	219	168	
Gender (Male)	108 (.49)	94 (.56)	
Ethnicity			
African American	26 (.12)	12 (.07)	
Hispanic	156 (.71)	130 (.77)	
White	24 (.11)	17 (.10)	
Other	10 (.05)	6 (.04)	
Missing	3 (.01)	3 (.02)	
ELL	110 (.50)	93 (.55)	
CELDT score Mean (SD)	3.05 (1.03)	3.0 (.90)	
	Mean (SD)	Mean (SD)	*p*-value
CST	319.87 (48.5)	334.93 (58.7)	
Outcome Measures			
3rd grade			
ORF	76.79 (26.1)	81.89 (29.6)	
Spelling	10.16 (4.6)	10.68 (4.9)	
PPVT ss	86.2 (11.4)	87.95 (10.8)	
WRMT			
WID ss	102.6 (8.4)	104.8 (8.9)	
WATT ss	103.58 (12.6)	106.93 (13.8)	
PC ss	100.93 (8.1)	102.45 (8.2)	
WC raw	17.01 (5.8)	17.61 (6.4)	
Predictor Measures			
1st grade			
TOLDrvss	7.3 (3.6)	8.3 (3.0)	.008
WIF	48.85 (22.5)	52.75 (24.5)	
ORF	52.37 (26.3)	55.04 (28.0)	
2nd grade			
TOLDrvss	7.76 (2.7)	8.2 (2.8)	
ORF	89.38 (28.19)	80.52 (31.9)	
WRMT			
WID ss	107.31 (11.8)	109.46 (11.0)	
WATT ss	105.63 (11.9)	110.31 (11.3)	<.001
PC ss	100.95 (9.6)	104.99 (9.0)	<.001
WC raw	10.75 (6.3)	9.64 (7.0)	

Note. p-values are reported for significant differences in means between cohorts; ELL= English Language Learner; CELDT= California English Language Development Test (range 1-5); PPVTss= Peabody

Picture Vocabulary Test standard scored (normed M=100; SD=15); TOLDrvss= Test of Language Development Relational Vocabulary Standard Score, 3rd Ed. (normed M=10; SD=3); ss standard Score

Table 4
CST ELA Score Descriptions

	Far Below Basic	Below Basic	Basic	Proficient	Advanced
3rd Grade	150-258	259-299	300-349	350- 401	402-600

Notes:
Advanced: This level represents a superior performance. Students demonstrate a comprehensive and complex understanding of the knowledge and skills measured by this assessment, at this grade, in this content area.
Proficient: This level represents a solid performance. Students demonstrate a competent and adequate understanding of the knowledge and skills measured by this assessment, at this grade, in this content area.
Basic: This level represents a limited performance. Students demonstrate a partial and rudimentary understanding of the knowledge and skills measured by this assessment, at this grade, in this content area.
Far Below / Below Basic: This level represents a serious lack of performance. Students demonstrate little or a flawed understanding of the knowledge and skills measured by this assessment, at this grade, in this content area

Table 5
Rotated Factor Solution Pattern Matrix for 3rd grade reading outcomes

	Factor 1 Word Reading/Fluency		Factor 2 Comprehension/Vocabulary	
	Cohort A	Cohort B	Cohort A	Cohort B
ORF	.783	.901		
WRMT WID	.822	.873		
WRMT WATT	.764	.762		
SPELLING	.912	.789		
WRMT WC			.516	.840
WRMT PC			.610	.376
PPVT			.434	.553

Note. ORF= Oral Reading Fluency; WRMT= Woodcock Reading Mastery Tests; WID= Word Identification; WATT= Word Attack; WC=Word Comprehension; PC Passage Comprehension; Spelling = Test of Spelling Development; PPVT= Peabody Picture Vocabulary Test

Table 6
Prevalence of RD readers by Method

	1.5 SD		1.0 SD	
	Within RD group	Within sample	Within RD group	Within sample
Cohort A				
Word Reading/ Fluency	11.8%	0.9%	31.1%	6.4%
Compre-hension/ Vocabulary	58.8%	4.6%	35.5%	7.3%
Both	29.4%	2.3%	33.3%	6.8%
TOTAL		**7.8%**		**20.5%**
Cohort B				
Word Reading/ Fluency			22.2%	4.8%
Compre-hension/ Vocabulary			27.8%	6.0%
Both			50%	10.7%
TOTAL				**21.5%**

Table 7

Univariate Means, p-values and effect sizes under Method1 (Cohort A)

	Word Reading/Fluency Composite				Comprehension/Vocabulary Composite			
	RD	Proficient			RD	Proficient		
N	7	212			15	204		
Prevalence	3.2%	96.8%			6.8%	93.6%		
	Mean (*SE*)	Mean (*SE*)	Difference (pvalue)	Effect Size (n^2)/ Cohen's *F*	Mean (*SE*)	Mean (*SE*)	Difference (*p*-value)	Effect Size (n^2)/ Cohen's *F*
ORF	24.71 (8.8)	78.02 (1.7)	**53.3 (*p*<.001)**	.15/.42	50.64 (6.5)	78.07 (1.8)	27.43 (*p*<.001)	.076/.29
SPELLING	2.29 (1.6)	10.23 (0.3)	**7.9 (*p* <.001)**	.103/.32	5.14 (1.17)	10.32 (0.3)	5.17 (*p*<.001)	.084/.30
PPVT0910 W	80.29 (4.3)	86.34 (0.8)	6.1 (*p*=.17)	.009/.1	69.64 (2.8)	87.36 (0.8)	**17.72 (*p*<.001)**	.154/.43
WRMT								
WID SS	87.71 (2.9)	102.84 (0.5)	**15.2 (*p*<.001)**	.112/.36	94.43 (2.1)	102.96 (0.6)	8.53 (*p*<.001)	.072/.28
WATT SS	88.14 (4.3)	103.25 (0.8)	**15.1 (*p*=.001)**	.057/.25	92.93 (3.0)	103.46 (0.8)	10.53 (*p*<.001)	.053/.24
PC SS	84.85 (2.8)	101.31 (0.5)	16.5 (*p*<.001)	.145/.41	89.07 (2.0)	101.62 (0.5)	**12.54 (*p*<.001)**	.162/.44
WC Raw	7.29 (2.1)	17.21 (0.4)	9.9 (*p*<.001)	.102/.34	8.5 (1.4)	17.49 (0.4)	**8.99 (*p*<.001)**	.161/.44
CST	261.43 (17.9)	321.97 (3.4)	6.55 (*p*=.001)	.053/.24	255.57 (12.1)	324.68 (3.3)	69.11 (*p*<.001)	.132/.39

Note. Bolded values are for construct-relevant measures; WRMT= Woodcock Reading Mastery Test; WID= Word Identification; WATT= Word Attack; PC=Passage Comprehension; WC=Word Comprehension; SS=Standard Score; SD=Standard Deviation. Published Normative WID and WATT Mean (SD) = 100(15); PPVT= Peabody Picture Vocabulary Test; W=winter; CST=California Standards Test.

Table 8

Univariate Means, p-values, and effect sizes under Method 2 (Cohort A)

	Word Reading/Fluency Composite				Comprehension/Vocabulary Composite			
	RD	Proficient			RD	Proficient		
N	29	190			31	188		
Prevalence	13.2%	86.7%			14.2%	85.8%		
	Mean (*SE*)	Mean (*SE*)	Difference (*p*-value)	Effect Size (n^2)/ Cohen's *F*	Mean (*SE*)	Mean (*SE*)	Difference (*p*-value)	Effect Size (n^2)/ Cohen's *F*
ORF	45.96 (4.3)	80.84 (1.7)	**34.88 (p<.001)**	.222/.53	55.4 (4.4)	79.8 (1.8)	24.2 (p<.001)	.131/.39
SPELLING	4.33 (0.77)	10.83 (0.3)	**6.49 (p<.001)**	.237/.56	6.3 (0.8)	10.57 (0.33)	4.26 (p<.001)	.11/.35
PPVTW	80.96 (2.2)	86.93 (0.86)	5.97 (p=.012)	.031/.18	72.62 (1.9)	88.41 (0.8)	**15.79 (p<.001)**	.233/.55
WRMT								
WID SS	92.63 (1.4)	103.87 (0.54)	**11.24 (p<.001)**	.225/.54	96.55 (1.4)	103.34 (0.59)	6.79 (p<.001)	.087/.31
WATT SS	90.44 (2.0)	104.63 (0.81)	**14.19 (p<.001)**	.173/.46	95.86 (2.11)	103.88 (0.87)	8.072 (p=.001)	.06/.25
PC SS	90.85 (1.34)	102.28 (0.53)	11.42 (p<.001)	.242/.56	92.69 (1.3)	102.1 (0.55)	**9.41 (p<.001)**	.17/.45

WC Raw	11.63 (1.03)	17.67 (0.41)	6.04 (p<.001)	.131/.39	9.66 (0.91)	18.08 (0.77)	**8.42 (p<.001)**	.269/.61
CST 3rd grade	274.96 (8.72)	326.83 (3.43)	51.87 (p<.001)	.133/.39	267.9 (8.1)	328.63 (3.3)	60.73 (p<.001)	.194/.49

Notes: Bolded values are for construct-relevant measures; WRMT=Woodcock Reading Mastery Test; WID= Word Identification; WATT= Word Attack; PCOMP=Passage Comprehension; WCOMP=Word Comprehension; SS=Standard Score; SD=Standard Deviation. Published Normative WID and WATT Mean (SD) = 100(15); PPVT= Peabody Picture Vocabulary Test; W=winter; CST=California Standards Test. Cohen's F statistic adjusts for the upward bias in Eta-Squared. Cohen's F= $\sqrt{\frac{n^2}{1-n^2}}$

Table 9

Univariate Means, p-values, and effect sizes under Method 2 (Cohort B)

	Word Reading/Fluency Composite				Comprehension/Vocabulary Composite			
	RD	Proficient			RD	Proficient		
N Prevalence	26	142			28	140		
	Mean (*SE*)	Mean (*SE*)	Difference (*p*-value)	Effect Size (n^2)/ Cohen's *F*	Mean (*SE*)	Mean (*SE*)	Difference (*p*-value)	Effect Size (n^2)/ Cohen's *F*
ORF	43.15 (4.8)	89.36 (2.0)	**46.21 (p<.001)**	.322/.69	51.79 (5.0)	87.9 (2.2)	36.11 (p<.001)	.208/.51
SPELLING	4.11 (0.77)	11.97 (0.33)	**7.86 (p<.001)**	.350/.73	5.3 (0.8)	11.8 (0.36)	6.5 (p<.001)	.246/.57
PPVTW	79.85 (2.0)	89.6 (0.85)	9.75 (p<.001)	.112/.36	76.79 (1.8)	90.18 (0.8)	**13.39 (p<.001)**	.221/.53
WRMT								

WID SS	93.12 (1.4)	107.09 (0.62)	**13.97 (p<.001)**	.325/.69	94.68 (1.5)	106.83 (0.65)	12.15 (p<.001)	.258/.59
WATT SS	92.65 (2.4)	109.74 (1.1)	**17.09 (p<.001)**	.201/.5	94.1 (2.4)	109.51 (1.1)	15.14 (p<.001)	.174/.46
PC SS	92.12 (1.34)	104.48 (0.58)	12.36 (p<.001)	.305/.66	91.43 (1.2)	104.7 (0.5)	**13.27 (p<.001)**	.367/.76
WC Raw	9.81 (1.1)	19.1 (0.46)	9.29 (p<.001)	.281/.63	8.7 (.94)	19.4 (.42)	**10.7 (p<.001)**	.397/.81
CST 3rd grade	268.42 (12.2)	345.12 (4.7)	76.7 (p<.001)	.19/.48	270.25 (11.9)	344.81 (4.6)	74.56 (p<.001)	.187/.48

Notes: Bolded values are for construct-relevant measures; WRMT=Woodcock Reading Mastery Test; WID= Word Identification; WATT= Word Attack; PCOMP=Passage Comprehension; WCOMP=Word Comprehension; SS=Standard Score; SD=Standard Deviation. Published Normative WID and WATT Mean (SD) = 100(15); PPVT= Peabody Picture Vocabulary Test; W=winter; CST=California Standards Test. Cohen's F statistic adjusts for the upward bias in Eta-Squared. Cohen's $F=\sqrt{\frac{n^2}{1-n^2}}$

Table 10
Pearson Correlations for 3rd grade outcomes

	Cohort B							
	ORF	WRMT WID	WRMT WATT	SPELLING	PPVT	WRMT PComp	WRMT WComp	CST
Cohort A								
ORF	1	.799	.639	.623	.353	.667	.544	.649
WRMT WID	.747	1	.746	.702	.492	.735	.636	.662
WRMT WATT	.575	.722	1	.672	.344	.660	.557	.536
SPELLING	.685	.754	.710	1	.359	.666	.468	.610
PPVT	.185	.246	.230	.174	1	.453	.468	.513
WRMT PComp	.602	.717	.594	.585	.298	1	.645	.662
WRMT WComp	.375	.453	.400	.448	. 304	.538	1	.579
CST	.560	.572	.449	.546	.459	.603	.540	1

Note: Cohort A in rows and below the diagonal, Cohort B in columns and above the diagonal; all coefficients are significant at p<.001; all scores are raw scores.

Table 11

Pearson Correlations for 1st and 2nd grade Predictors and 3rd grade Outcomes

	1st WIF	1st ORF	2nd ORF	2nd WID	2nd WATT	2nd WC	2nd PC	1st TOLD	2nd TOLD
1st WIF	1	.920	.846	.631	.650	.622	.631	.328	.367
1st ORF	.905	1	.865	.636	.687	.696	.689	.365	.404
2nd ORF	.789	.831	1	.672	.689	.626	.675	.288	.385
2nd WID	.541	.607	.549	1	.804	.554	.832	.447	.405
2nd WATT	.573	.645	.518	.827	1	.551	.747	.302	.370
2nd WC	.598	.655	.567	.510	.576	1	.674	.505	.475
2nd PC	.570	.623	.554	.845	.783	.623	1	.463	.506
1st TOLD	**.140**	.208	.125(ns)	.252	.249	.247	.327	1	.545
2nd TOLD	.260	.331	.256	.372	.431	.426	.529	.454	1
Cohort A									
3rd ORF	.805	.827	.882	.562	.540	.546	.565	**.159**	.263
3rd WID	.637	.689	.657	.847	.800	.568	.793	.251	.344
3rd WATT	.574	.613	.508	.713	.783	.542	.453	.237	.339
3rd Spell	.691	.704	.679	.632	.689	.649	.657	.246	.361
3rd WC	.369	.375	.364	.350	.387	.535	.453	.357	.460

3rd PC	.508	.555	.533	.760	.687	.520	.780	.265	.428
3rd PPVT	**.151**	.245	.247	.237	.280	.336	.372	.397	.506
Cohort B									
3rd ORF	.797	.820	.9	.583	.609	.584	.599	.283	.388
3rd WID	.646	.674	.732	.679	.696	.608	.726	.411	.473
3rd WATT	.585	.607	.628	.675	.746	.503	.617	.266	.339
3rd Spell	.652	.695	.694	.614	.705	.600	.615	.375	.376
3rd WC	.442	.485	.537	.452	.487	.597	.596	.332	.493
3rd PC	.587	.605	.651	.753	.664	.609	.756	.406	.451
3rd PPVT	.226	.302	.318	.348	.296	.391	.427	.410	.409

Note. Top section: Cohort A correlations on rows and below diagonal; Cohort B correlations on columns and above diagonal; Bottom section: Correlations of predictors with outcomes by cohort; all ps<.01 except bolded = p<.05 and ns=not significant.

Table 12

Logistic Regression Models for Word Reading/Fluency outcomes under Method 2 (Cohort A)

	b(SE)	Odds Ratio	*Best Predictors*[a]	Fit Criteria					
				AIC		BIC		-2 loglikelihood	
				Step	*Best Predictors*	*Step*	*Best Predictors*	*Step*	*Best Predictors*
Null				11004.8		110048.74		5489.40	
Model 1									
Step 1				3671.12	2037.66	3701.54	2054.56	1826.37	1013.83
Gender	-1.86**	.156	-1.82***						
	(.580)		(.57)						
Hispanic	-.354	.702							
	(.505)								
WIF1s	-.071*	.932	-.106***						
	(.033)		(.02)						
ORF1s	-.033	.986							
	(.027)								
Step 2				8092.57	3592.12	8183.57	3622.54	4019.29	1787.06
Gender	-2.34**	.097	-2.30**						
	(.875)		(.855)						
WIF1s	-.001	.999							
	(.037)								
ORF2s	-.109***	.897	-.101***						
	(.028)		(.017)						
WID 2ss	.024	1.02							
	(.054)								
WATT 2ss	-.086	.917							
	(.049)								
PC2ss	-.158*	.854	-.21***						
	(.079)		(.052)						

Final Model			3592.12	3622.54	1787.06
Gender	-2.302** (.855)	.100			
ORF2s	-.101*** (.017)	.904			
PC2ss	-.209*** (.052)	.812			

Note. a= beta coefficients(se)of model with only significant predictors at each step; s=spring time point; ss=standard score;1=1st grade; 2=2nd grade; WIF=Word Identification Fluency; ORF=Oral Reading Fluency; WID= Word Identification; WATT= Word Attack; PC=Passage Comprehension; Final Model includes only significant predictors from steps 1 and 2 of Model 1.
**p<.05; **p<.01; ***p<.001*

Table 13

Logistic Regression Models for Comprehension/Vocabulary outcomes under Method 2 (Cohort A)

	b(SE)	Odds Ratio	*Best Predictors*[a]	Fit Criteria					
				AIC		BIC		-2 Loglikelihood	
				Step	*Best Predictors*	*Step*	*Best Predictors*	*Step*	*Best Predictors*
Null Model				11575.68		11633.14		5770.84	
Model 1									
Step 1				4853.69	1328.51	4907.77	1345.41	2410.85	659.23
Gender	-1.09*	.33	-.99*						
	(.45)		(.41)						
Hispanic	-.09	.92							
	(.57)								
ELL1	.21	1.23							
	(.54)								
TOLD1ss	-.17**	.84	-.18***						
	(.06)		(.05)						
ORF1	-.05	.96							
	(.03)								
WIF1	-.02	.98							
	(.03)								
Step 2				6979.105	3151.10	7070.36	3181.52	3462.55	1566.55
Gender	-1.12*	.33	-.92*						
	(.47)		(.47)						
TOLD1ss	-.06	.94							
	(.08)								
TOLD2ss	-.29*	.75	-.41***						
	(.12)		(.09)						
ORF2s	-.02*	.97	-.04***						
	(.01)		(.01)						

WC2raw	-.14 (.07)	.87			
PC2ss	.007 (.032)	.99			
Final Model			3151.10	3181.52	1566.55
Gender	-.923* (.47)	.40			
TOLD2ss	-.41*** (.09)	.67			
ORF2s	-.04*** (.011)	.96			

Note. a= beta coefficients(se)of model with only significant predictors at each step; s=spring time point; ss=standard score; raw=raw score (only raw scores were available for WC); 1=1st grade; 2=2nd grade; WIF=Word Identification Fluency; ORF=Oral Reading Fluency; WC=Word Comprehension; PC=Passage Comprehension; TOLD= Test or Oral Language Development; Final Model includes only significant predictors from steps 1 and 2 of Model 1.

Table 14
Classification Table for RD in Word Reading/Fluency by Cohort

	Sensitivity		Specificity	
	Cohort A	Cohort B	Cohort A	Cohort B
1st grade WIF				
25th %tile	75%	80.8%	81%	81.9%
33rd %tile	82.1%	88.5%	76.1%	76.1%
2nd grade Passage Comprehension				
25th %tile	79.3%	80%	81.6%	83.6%
33rd %tile	89.7%	88%	73.5%	74.3%
2nd grade ORF				
Published Final Benchmark (90wcpm)	100%	100%	55.9%	43.2%
Sample Final Benchmark (Cohort A: 75wcpm) (Cohort B: 70wcpm)	100%	100%	78.8%	69.1%
Low Growth (1 SD below sample mean)	37.9%	69.6%	89.4%	95%
Dual Discrepancy (<90wcpm and <1SD below mean growth	40.7%	69.6%	95.1%	95%
25th %tile	85.2%	84.6%	83.2%	85.6%
33rd %tile	100%	96.2%	75.5%	79.1%

Table 15
Classification Table for RD in Comprehension/Vocabulary by Cohort

	Sensitivity		Specificity	
	Cohort A	Cohort B	Cohort A	Cohort B
1st grade TOLD				
25th %tile	50%	46.4%	71.8%	79.4%
33rd %tile	56.3%	71.4%	66.3%	61%
2nd grade TOLD				
25th %tile	66.7%	53.6%	78.7%	83.6%
33rd %tile	72.7%	64.3%	66.9%	72.9%
2nd grade ORF				
Published Final Benchmark (90wcpm)	81.3%	96.4%	54.2%	43.9%
Low Growth (1 SD below sample mean)	24.2%	53.6%	87.5%	93.6%
Dual Discrepancy (<90wcpm and <1SD below mean growth	25%	53.6%	93.2%	93.6%

25th %tile	46.9%	71.4%	78.2%	83.5%
33rd %tile	59.4%	78.6%	70.4%	76.3%

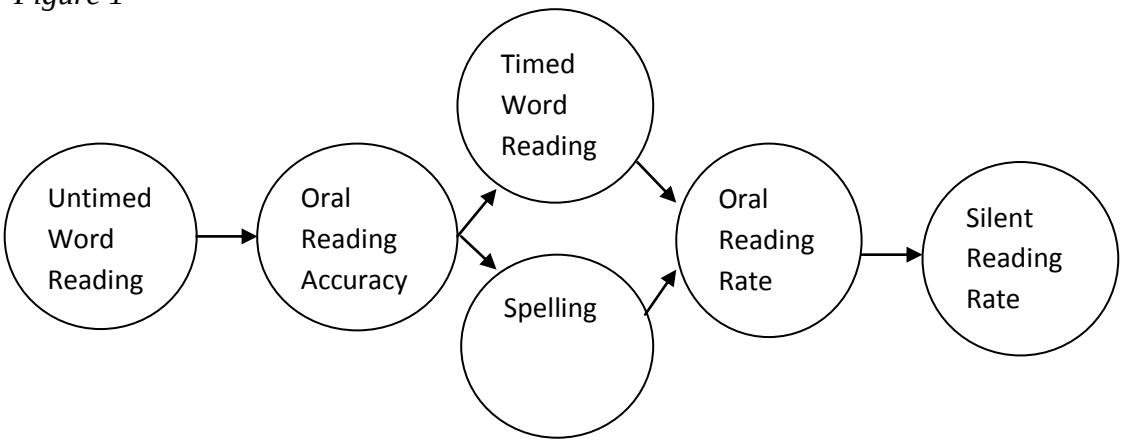
Table 16
Final Logistic Model Sensitivity and Specificity by Cohort

	Sensitivity		Specificity	
	Cohort A	Cohort B	Cohort A	Cohort B
Word Reading/Fluency	86.2%	80%	88.9%	89.2%
Comprehension/Vocabulary	84.4%	82.1%	76.7%	69.8%

Note. Sensitivity and specificity indices reported here are for final models for each of the WR-F and C-V logistic regression analyses

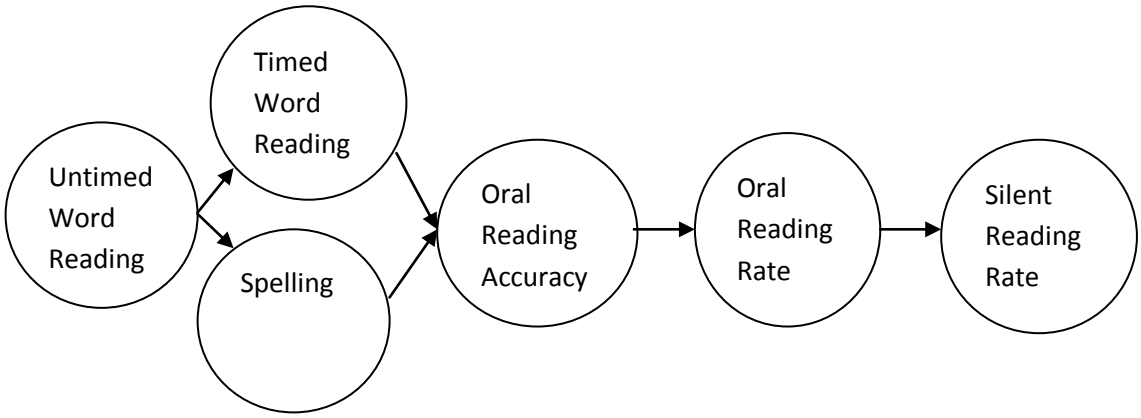
Figures

Figure 1



Caption: Figure adapted from Morris, D., Trathen,W., Lomax, R.G., Perney, J., Kucan, L., Frye, E. M., …Schlagal, R. (2012). Modeling aspects of print-processing skill: implications for reading assessment. Reading and Writing, 25(1), 189-215.

Figure 2



Caption: Figure adapted from Morris, D., Trathen,W., Lomax, R.G., Perney, J., Kucan, L., Frye, E. M., …Schlagal, R. (2012). Modeling aspects of print-processing skill: implications for reading assessment. Reading and Writing, 25(1), 189-215.