Relationship of Oral Reading Fluency Probes on Students' Reading Achievement Test Scores

by

Sarah Devena

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2013 by the
Graduate Supervisory Committee:

Linda C. Caterino, Chair
John Balles
Sarup Mathur

ARIZONA STATE UNIVERSITY

December 2013

Abstract

Current emphasis on adequate academic progress monitored by standardized assessments has increased focus on student acquisition of required skills. Reading ability can be assessed through student achievement on Oral Reading Fluency (ORF) measures. This study investigated the effectiveness of using ORF measures to predict achievement on high stakes tests. Study participants included 312 students across four Title 1 elementary schools in a Southwestern United States school district utilizing the Response to Intervention (RTI) model. Participants' ORF scores from first through third grade years and their third grade standardized achievement test scores were collected. In addition, information regarding reading interventions was obtained. Pearson product-moment correlations were used to determine how ORF scores and specific reading skills were related. Correlations were also used to assess whether the ORF scores from the fall, winter, or spring were most related to high stakes test scores. Additionally, the difference between computer-based versus instructor-led interventions on predicting high stakes test scores was assessed. Results indicated that correlation coefficients were larger between ORF and reading comprehension scores than between ORF and basic reading skills. ORF scores from spring were more highly related to high stakes tests than other times of the year. Students' ORF scores were more strongly related to high stakes tests when in computer-based interventions compared to instructor-led interventions. In predicting third grade high stakes test scores, first grade ORF scores had the most variance for the non-intervention sample, while third grade ORF scores had the most variance for the intervention sample.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

Educational Law Background

Education and law have been closely linked over the last half century in America. Various federal laws from 1965 to the present have been enacted which have had a substantial impact on state educational policies (Yell & Dragow, 2007). On April 9, 1965 Congress enacted the Elementary and Secondary Education Act (ESEA) of 1965 (PL 89-10), which outlined standards of education, along with the federal funding for educational programs. Although preliminary educational policy laws for students with disabilities had begun to emerge, the Education for All Handicapped Children Act (EAHCA) of 1975 (PL 94-142) was the first law enacted that mandated a free, appropriate public education for all students. This law was renamed and updated to the Individuals with Disabilities Education Act (IDEA) of 1990 (PL 101-476). The 1997 IDEA Amendments (PL 105-17), reauthorized IDEA with updates on special education standards. In 2001, ESEA was reauthorized and named the No Child Left Behind Act (NCLB) of 2001 (PL 107-110). This law served to outline the updated federal regulations on curriculum and testing standards for all school children. IDEA was reauthorized and amended in 2004 to become the Individuals with Disabilities Education Improvement Act (IDEIA) of 2004 (PL 108-446), which continued implementing regulations for children with special needs (U.S. Department of Education, 2006).

Student reading acquisition has been a primary focus in more recent educational law. In 1997, Congress mandated the National Research Council to research reading acquisition. The council published its findings in, *Preventing Reading Difficulties in*

Young Children (1998), which indicated that reading, or deriving meaning from written text, was based on five basic skills: Phonemic awareness, phonics, fluency, comprehension, and vocabulary. Following this publication, the National Reading Panel (2000) was formed, which published a report emphasizing explicit and systematic instruction of these five essential components of literacy.

The No Child Left Behind Act of 2001 mandated that in order to receive federal funding, each school must make adequate yearly progress in both reading and mathematics, as evidenced by scores on state tests. Proficiency is expected for all students by 2014. The NCLB law encourages use of evidence-based instruction to increase student success. Each state is responsible for its own curriculum standards and assessment measures to meet the NCLB standards (Stansfield, 2011).

The reports and findings from the National Research Council (1998), along with the National Reading Panel (2000), led to the enactment of the national reading initiative component of NCLB entitled Reading First. This program provided assistance to states to establish scientifically-based reading programs for children, as well as tools for professional development, instruction, and assessment. Suggested educational assessments included information gathered through screening, diagnostic, and instructional procedures.

At the state level, Arizona formed the Arizona State Reading Task Force in 1998, which focused on developing a research-based reading curriculum. The task force recommended that each school district develop measures to assess reading proficiency for students in kindergarten through third grade (Arizona Reading Success Task Force,

2000). These measures focused on four of the five components of reading developed by the National Reading Panel (2001): Phonemic awareness, the alphabetic principle, comprehension, and basic reading skills. Subsequently, the comprehensive reading initiative, Arizona READS (2001), designed to improve reading achievement was enacted. Arizona READS states that every student should gain reading proficiency by third grade and remain proficient through twelfth grade. Starting in 2014, students who do not reach the established proficiency standards by third grade will be retained.

Reading achievement is prioritized in the Arizona Department of Education (ADE) school accountability system, Arizona LEARNS (2001). Arizona LEARNS requires that Arizona schools conduct ongoing measurement of student skill acquisition. In the second grade, students in Arizona schools take the Stanford Achievement Test-10th Edition (SAT-10; Harcourt Educational Measurement, 2003a). The Arizona Department of Education developed the Arizona Instrument to Measure Standards (AIMS; Arizona Department of Education, 2012a) assessment for students at or above a third grade placement to assess state content standards. Questions developed for the AIMS assessment are combined with the SAT-10, which forms the AIMS Dual Purpose Assessment (DPA). Some Arizona schools also adopted curriculum measurement systems including the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002) and the System to Enhance Educational Performance (STEEP; Witt, 2007) to help monitor children's acquisition of essential reading components.

Both the DIBELS and STEEP systems are standardized assessment systems used to screen and monitor specific educational skills. Along with other reading measures, the

DIBELS and STEEP systems both include one minute reading fluency measures. These screening measures are typically given individually to students three times per year. Subsequently, students are grouped in levels of proficiency according to benchmark scores provided by the STEEP and DIBELS systems. The lower performing students are given targeted interventions and are often monitored weekly to determine intervention success (Good & Kaminski, 2002; Witt, 2007).

These two assessment systems, although similar, have distinct differences. The DIBELS system includes assessment in Letter Naming Fluency, Initial Sound Fluency, Phoneme Segmentation Fluency, Nonsense Word Fluency, Retell Fluency, Oral Reading Fluency, and Word Use Fluency, dependent upon student grade level. The STEEP system includes Oral Reading Fluency and Sentence Maze Fluency. The DIBELS system groups children for the beginning, middle, and end of the year in the categories of at-risk, some risk, or low risk. The STEEP system uses the end of the year proficiency levels of frustrational, instructional, and mastery. In addition to reading skills, the STEEP system also includes measures in writing and math.

Response to Intervention Model

Prior to IDEIA, the general concept presented by court law and the United States Office of Education was that students needed an IQ-achievement discrepancy to be diagnosed with a learning disability (LD; Herr, 2003). However, IDEIA recognized that the use of alternative methods of assessment were acceptable for a diagnosis of LD including the student's response to scientifically-based interventions (known as Response

to Intervention or RTI) and specifically prohibited states from *requiring* this discrepancy model, although it was still offered as an alternative.

Generally, RTI is a leveled system founded on intervention and prevention; however, since many procedures in the RTI process were not specified under IDEIA, it is implemented in different ways across school districts (Fuchs, Fuchs, & Compton, 2012). The RTI model incorporates intervention in a multitier, or multilevel, system of supports. This system uses a methodized process to systematically analyze data to identify and target academic difficulties (Brown-Chidsey & Steege, 2006). Generally, this process involves giving students effective instruction, monitoring their progress, providing interventions to those falling behind, continuing to monitor progress, and providing a higher level of support for students when necessary (Brown-Chidsey & Steege).

Alternative RTI models vary in the number of intervention tiers, or levels, offered. In general, the levels indicate changes in the intervention intensity, group size, or leader qualifications (Fuchs, Mock, Morgan, & Young, 2003). Currently, Fuchs et al. (2012) describe three levels of supports including primary, secondary, and tertiary prevention. A higher prevention level denotes stronger research support for interventions and better instructor qualifications. During the primary prevention level students receive support through general education instructional practices including core curriculum and classroom routines. At this primary level, students are screened to determine those at risk for not responding to the primary level supports. In the secondary prevention level, students are given empirically validated small group interventions. Students are assessed at this secondary level to determine their response to intervention and need for movement

between levels. Tertiary prevention is offered to students not responding to the first two levels of support. This prevention level is given by instructional specialists knowledgeable in effective curriculum for unique learners. Frequent data is collected to determine student response to instruction. In the Fuchs et al. (2012) model the tertiary level is specified as special education services. Alternative perspectives on RTI do not include special education services within the RTI framework (Brown-Chidsey & Steege, 2006).

A foundational concept across RTI models is that effective instruction must be implemented for all students (Brown-Chidsey & Steege, 2006). Instruction in reading has been described by the National Reading Panel report (2000). The comprehensive meta-analysis focused on appropriate instruction by examining available research studies using controlled experiments with measurable results published in peer reviewed journals. Results were positive for phonemic awareness, phonics, fluency, comprehension, and vocabulary instruction.

Phonemic awareness instruction is designed to teach skills associated with the manipulation of the smallest part of the spoken language called phonemes (National Reading Panel, 2000). According to the National Reading Panel meta-analysis, training in phonemic awareness led to significant gains across grades in reading and spelling skills. Many methods for teaching phonemic awareness were successful, but these methods needed to be tailored to student needs. Phonics instruction is aimed at relating letters to sounds (phonemes) in order to read (National Reading Panel). Instruction can be delivered by using sounds to build words or determining the sounds in whole words.

Research reviewed by the National Reading Panel supported teaching sequential phonics concepts systematically for children in kindergarten through sixth grade to help increase decoding and word recognition skills (National Reading Panel). Fluency instruction is aimed at increasing the ability to read aloud with appropriate speed, precision, and intonation (National Reading Panel). Research findings from the meta-analysis suggested guided oral reading significantly increased fluency as well as word recognition and comprehension (National Reading Panel). Comprehension instruction is aimed at the intentional process of relating to a sequence of text (National Reading Panel). Research reviewed by the National Reading Panel indicated that teaching various techniques, or strategies, was most effective in recalling, questioning, and summarizing texts. Vocabulary instruction is targeted at oral and written word enhancement (National Reading Panel). The National Reading Panel meta-analysis indicated that vocabulary development was increased with multiple exposure, varied teacher methods as well as contexts, computer enhanced learning, and task restructuring.

After students have been identified for RTI interventions and these interventions have been administered, their response to these interventions must be carefully monitored. Measurement of progress is essential to the RTI approach. Progress measures need to be sensitive to growth as a student's skills improve (Marston, 1989). Curriculum-Based Measurement (CBM) is an appropriate tool which can be used to measure the effectiveness of instruction on student learning because it is sensitive enough to detect small changes in performance (Hintze, Christ, & Methe, 2006). CBM is a brief, standardized measure of academic skills (Shinn, 1995) first developed to test the

effectiveness of program modification (Deno & Mirkin, 1977). Within the RTI approach, CBM is used for problem identification, instructional placement, goal-setting, intervention planning, progress monitoring, and eligibility decisions (Fuchs & Fuchs, 1997).

Due to the emphasis on adequate progress monitored by standardized assessments, many schools use CBMs to identify students at risk for failing these summative tests: CBMs have been developed to measure different academic competencies in reading, math, and writing. Reading CBMs can range from requiring a student to read individual letter sounds, to reading consonant vowel consonant (CVC) words, to reading longer words, to reading full sentences or even paragraphs, as well as reading comprehension skills. CBMs are usually timed, and the number of correct and incorrect letter or word identifications can be graphed to measure progress over time. Shinn (2007) recommends that data be collected one or two times per week and intervention success be determined after 7 to 10 data points.

According to Adams (1990), the strongest component of skillful reading is the speed at which someone can read a text and reproduce it into spoken language, referred to as oral reading fluency (ORF). ORF is essential because it measures perceptual skills, the ability to convert letters into sound representations, and the comprehension of meaningful connections within and between sentences (Fuchs, Fuchs, Hosp, & Jenkins, 2001). Some experts consider reading fluency to be "one of the defining characteristics of good readers, and a lack of fluency is a common characteristic of poor readers" (Hudson, Lane, & Pullen, 2005, p. 702). There is strong research support that reading fluency accurately

measures reading comprehension (Reschly, Busch, Betts, Deno, & Long, 2009; Shinn, Good, Knutson, Tilly, & Collins, 1992). Fuchs, Fuchs, and Maxwell (1988) assessed the correlation of direct measures of reading comprehension, such as passage recall or question answering, and the indirect measure of ORF. Results indicated that the correlation between ORF and reading comprehension was significantly higher than any of the direct measures of reading comprehension. The correlation between ORF and reading comprehension is generally strongest in elementary school and decreases as students age (Fuchs et al., 2001).

ORF probes, the most commonly used reading CBMs (Stecker, Fuchs, & Fuchs, 2005), are used as components in the RTI model to enable interventions with specific remediation strategies in reading instruction (Coulter, Shavin & Gichuru, 2009). The effectiveness of using these reading measures to predict achievement is vital in the RTI process (Wood, 2006).

Reading Interventions

Within the RTI model, levels of intervention are targeted at the students who do not respond to effective instructional techniques. Extensive research has been completed on reading intervention efficacy (Goodwin & Ahn, 2010; Yang, 2006). In order to determine the effects of interventions on increasing reading skills, Yang completed a comprehensive meta-analysis which included 39 eligible experimental or quasi-experimental studies which focused on reading fluency in elementary students. Outcomes were assessed for speed, accuracy, and comprehension. Results indicated mean effect sizes in the medium range across interventions. For specific intervention

types, reading fluency skills were best taught using methods involving guided and repeated reading of connected texts. Additionally, interventions with repetitive practice had stronger treatment effects than those without repetitive practice across the outcome measures of rate, accuracy and comprehension. The participants needing remedial instruction or having disabilities had a larger mean effect size across studies than typical students did. No significant differences were found in the intervention efficacy across grade levels. For reading comprehension outcomes, fluency and comprehension strategies worked better than fluency-only interventions (Yang).

In order to assess morphological interventions on struggling readers, Goodwin and Ahn (2010) completed a comprehensive meta-analysis with 17 experimental studies. Morphological interventions are aimed at increasing understanding and manipulating the structural units, or morphemes, of words. Positive treatment effects were found for phonological awareness (d = 0.49), morphological awareness (d = 0.40), vocabulary (d = 0.40), reading comprehension (d = 0.24), and spelling (d = 0.20). Students with speech and language delays showed the strongest treatment effects (d = 0.77), followed by English language learners (d = 0.62), struggling readers (d = 0.46), students with learning disabilities (d = 0.22), and lastly students with reading disabilities (d = 0.17). Results indicated that instruction in morphemes may be critical for interventions targeted at students with language deficits.

Other studies have shown that small group reading and remediation interventions can efficaciously target deficits in basic reading skills (Fuchs et al., 2008; Ryder, Tunmer, & Greaney, 2008; Tucker & Jones, 2010). Short term, intensive interventions

aimed at elementary students showed positive gains over control groups in reading fluency, rate, accuracy (Tucker & Jones, 2010) phonemic awareness, decoding skills, and word reading (Fuchs et al., 2008; Ryder, Tunmer, & Greaney, 2008).

Research has generally supported the use of computer-based interventions in improving reading skills (Blok, Oostdam, Otter, & Overmaat, 2002; Gatti Evaluation, 2011; Layng, Twyman, & Stikeleather, 2003; National Reading Panel, 2000; Regtvoort & van der Leij, 2007; Saine, Lerkkanen, Ahonen, Tolvanen, & Lyytinen, 2011; Tamim, Bernard, Borokhovski, Abrami, & Schmid, 2011; Torgesen, Wagner, Rashotte, Herron, & Lindamood, 2010). Based on 21 studies meeting the inclusion criteria, the National Reading Panel study (2000) indicated that all studies showed positive results. The first promising area included using speech to supplement computer text. The second area of promise included using hypertext that relates text to supporting materials. The third area included word processing techniques in order to integrating writing into instruction. However, research was lacking in specific instructional applications.

Following the National Reading Panel (2000) meta-analysis, Blok et al. (2002) assessed the influence of specific types of computerized instruction on reading achievements. The instructional categories included: (a) phonological awareness, (b) word reading with speech feedback, (c) time limit exposure word reading, (d) text reading with speech feedback, (e) reading/listening, and (f) mixed methods. Their meta-analysis included 42 published studies with the average participant age of 8.5 years. The research findings for computer based instruction indicated an overall small effect size (d = 0.20). The two main moderating variables were higher pretest scores and English as

the language of instruction. These two variables related to higher effect sizes. Interestingly, the type of instructional category had no impact on study outcomes. Retention of skills yielded mixed results with only some study samples keeping treatment effects in follow up analyses (Blok et al.).

Tamim et al. (2011) used a second-order meta-analysis to review research conducted on the use of technology to supplement instruction. Inclusion results yielded 25 meta-analyses involving 1,055 initial studies from 1985 to present. Effect sizes across studies ranged from low to moderate. Effect sizes were stronger in kindergarten through twelfth grade and lower in postsecondary school (d = 0.40 vs. d = 0.29 respectively). In addition, Effect sizes were stronger when technology was used to support learning in place of direct instruction (d = 0.42 vs. d = 0.31 respectively). Results indicated that technology showed positive effects, but additional research should be completed on the specific points of interest (Tamim et al.).

Computer-based interventions have consistently shown reading gains in individual experimental studies. Torgesen et al. (2010) completed a study comparing two types of computer-based interventions. Participants included 112 first graders across three elementary schools. Three groups were formed with two intervention groups receiving computer-based supplemental interventions. One of the computer-based interventions focused on directly teaching phonemic spelling, whereas the other program stressed oral motor skills. The students were exposed to 50 minutes of interventions four times per week over the course of a school year. Outcome measures did not vary as a result of type of computer-based intervention. All students in the computer based

interventions had increased performance on phonological awareness, rapid naming, phonemic decoding, word reading accuracy/fluency, spelling, and reading comprehension by the end of their first grade year. At the one-year follow-up students receiving computer-based interventions had better outcomes than the control group in phonemic decoding, rapid naming, and spelling (Torgesen et al.).

Regtvoort and van der Leij (2007) assessed the effects of computer based interventions for kindergarten aged students at risk for dyslexia. Participants were divided into three groups including a control group with typical children (n = 16), at-risk children not receiving interventions (n = 26), and at-risk students receiving the interventions (n = 31). The computer-based interventions included training in word building specifically the pre-reading skills of letter-sound correspondence and phonemic awareness. Children received the interventions from their parents 10 minutes per day, 5 times per week for 10 weeks. Children exposed to the computer-based intervention made significant gains in phonemic awareness and letter knowledge skills. However, treatment effects did not last through their first or second grade years (Regtvoort & van der Leij).

Although both small group and computer-based interventions have shown efficacy, information on the comparison between the two types is critical for decision making. Saine et al. (2011) explored computer-based intervention versus small group remedial instruction for at-risk readers. The participants included 166 seven-year-old Finnish students. The participants were screened for pre-reading skills, then based on these results divided into three groups including remedial instruction (n = 25), computer-based instruction (n = 25), and mainstream support (n = 116). The longitudinal study

followed the children from first through third grade to determine the long-term intervention effects. Both the remedial and computer-based groups were taught a phonics-based program in a small group format. The participants were given 45 minutes of reading interventions four times per week for 28 weeks. While the children in the remedial program were completing 15 minutes of reading activities the computer based participants were exposed to a computer application for drill and practice of pre-reading and reading skills for the same amount of time. Children in both the remedial and computer-based interventions groups made significant gains in letter knowledge, decoding, and accuracy; however, children in the computer groups also made significant gains in fluency and spelling skills (Saine et al.).

One of the main limitations to generalizing Saine et al.'s (2011) findings is that all the children were Finnish speaking. It is not clear as to how these results apply to English speaking children. In addition, the limited sample size in each group makes replication necessary. Although this experiment indicates the addition of 15 minutes of computer-based reading time is beneficial, it is unclear if one intervention method over the other is superior.

In summary, research has supported the efficacy of small group interventions to target a wide range of reading deficits (Goodwin & Ahn, 2010; Fuchs et al., 2008; Ryder, Tunmer, & Greaney, 2008; Yang, 2006). Interventions including repeated practice showed the best outcome measures in rate, accuracy, and comprehension (Yang, 2006). Morphological awareness interventions showed the best efficacy with students struggling with language based deficits (Goodwin & Ahn, 2010). Computer-based interventions

have also proven efficacious in targeting reading deficits (Blok et al., 2002; National Reading Panel, 2000; Regtvoort & van der Leij, 2007; Tamim, et al., 2011; Torgesen et al., 2010). Computers have shown positive results for speech generated by computers, highlighting text, and the use of word processors (National Reading Panel, 2000). Generally it has been efficacious to use computers to supplement instruction rather than to replace direct instruction (Tamim et al.). However, research has not indicated significant differences in the type of computer program used (Blok et al.; Torgesen et al.). In addition, not all gains in performance and achievement had long term efficacy (Regtvoort & van der Leij; Torgsesesn et al.). Research on comparing instructor-led interventions to computer based interventions is limited; however, preliminary results indicated that the addition of computer based interventions may positively affect gains in fluency and spelling skills above what other remedial programs provide (Saine et al., 2011).

Reading Curriculum Based Measurement

There has been extensive research on the reliability and validity of reading CBMs. Early research on reading CBM use, as reviewed by Marston (1989), indicated strong support for using reading aloud proficiencies and word identification skills as indicators of global measures of reading ability. However, Marston's study used across-grade validity coefficients and multiple measures of reading ability, which makes interpretation difficult. Good and Jefferson (1998), updated the CBM meta-analysis to include those CBM studies that used within-grade validity coefficients, publically available criteria for validity coefficients (as opposed to teacher made tests or experimental measures), and

oral reading fluency. Validity coefficients were included for published norm-referenced tests, criterion-referenced tests, and basal reader series. Across the grade levels, all the reading coefficients ranged from .60 to .80 indicating strong support for the validity of the CBM construct (Good & Jefferson).

Following Marston (1989) and Good and Jefferson's (1998) meta-analyses, Wayman, Wallace, Wiley, Tichá, and Espin (2007) updated research on the technical adequacy of reading CBMs. Their research review included 64 articles on reading measures published from 1981 through 2005. The stability coefficients from the studies ranged from .84 to .99, and the equivalence coefficients ranged from .56 to .98. The meta-analysis also analyzed criterion validity evidence; specifically, the correlations between scores on general reading measures with scores on reading aloud, maze selection (a measure of reading comprehension), and word identification measures. Reading aloud measures were included in 84% of the studies, which included participants in first through eighth grades. The groups were diverse and included deaf or hard of hearing, English language learners, general education, high achieving, low achieving, special education, and visually impaired students, as well as non-differentiated groups (Wayman et al.). The correlation coefficients for the reading aloud measures and state reading achievement tests from Iowa, Michigan, Minnesota, and Washington ranged from .43 to .80. The correlation between reading aloud measures and norm-referenced reading tests ranged from .21 to .93. Reading sentences aloud was the best indicator of reading comprehension, even when compared to more typical measures of comprehension. Correlation coefficients for reading aloud and norm-referenced reading tests decreased as

students' grade levels increased. Interestingly, the CBM reading aloud measures had a tendency to overestimate the performance of African-American students, and to underestimate the performance of Caucasian-American students (Wayman et al., 2007).

A meta-analysis by Reschly et al. (2009) addressed the relationship of oral reading measures and other indices of reading ability, including state and national tests. The analysis included 41 studies from 1980 to 2008 with participants in the first through sixth grades. Participant groups included English language learners (ELL), students with free and reduced lunch, special education students (including those with learning disabilities), students receiving services through Chapter One funding, and students in regular education programs. Correlation of oral reading measures with national tests ranged from .35 to .91 (Reschly et al.). The median correlation coefficient had a weighted average of .67. ORF correlation coefficients with state tests from Arizona, Delaware, Colorado, Illinois, Ohio, and Oregon ranged from .43 to .81. Specifically, the Arizona state test and the oral reading fluency measure had a correlation coefficient of .74. Significant moderating factors from the analysis included: (a) Source of test (i.e. state specific versus nationally normed test), (b) group or individual administration, (c) length of time between CBM and criterion variable, and (d) reading subtest type (Comprehension, Vocabulary, Word Identification, Decoding, and Total Reading Score). Correlation coefficients were significantly higher for norm-referenced tests than for state specific tests. Reading passage CBMs were strongly associated with letter word tests, yielding higher correlation coefficients when individually administered. Correlation coefficients decreased with increased time between the probe and criterion variable.

Consistent with other students, ELL students had moderate to strong reliability and validity coefficients for reading aloud measures, as well as high stakes tests (Reschly et al., 2009).

Yeo (2010) conducted a meta-analysis assessing the predictive validity of ORF and state wide achievement tests. The analysis included 27 group design studies published from 2001 through 2007. Participants included students in first through eighth grade, with 11 of the studies only including third grade students. Group variables included free or reduced-price lunch, gender, English language learners, grade level, and special education status. Fourteen states were included in this analysis: Arizona, Colorado, Delaware, Florida, Illinois, Michigan, Minnesota, North Carolina, Ohio, Oklahoma, Oregon, Pennsylvania, Texas, and Washington. According to Cohen's (1992) criteria for correlation coefficients, the analysis revealed a large correlation coefficient (.69) across states. Results indicated that Arizona had a correlation coefficient of .74 between the ORF and the state achievement test. Primary moderators included personal characteristics, such as students' ELL or disabilities status and time of administration, with longer intervals between CBMs and standardized tests yielding lower correlation coefficients (Yeo).

Across studies, there was large variability between study correlation coefficients depending on study sample and measurement tool selection. For example, a sample from Florida had a correlation coefficient of .41 between the state total reading test and the ORF measure compared to .74 for an Arizona state reading test and ORF measure. On a national reading test, one study had correlation coefficients with ORF measures ranging

from .20 for comprehension up to .87 for vocabulary (Good & Jefferson, 1998; Reschly et al., 2009; Wayman et al., 2007; Yeo, 2010).

Stanford Achievement Test-10th Edition

A few studies have addressed the relationship between ORF probes and the SAT-10. Research commonalities include focusing on low-income students and utilizing the DIBELS ORF measure. Baker et al. (2008) conducted an analysis of the relationship between ORF and the SAT-10. The analysis included 34 Oregon Reading First funded schools in 16 school districts, with 8 districts located in large urban areas. Four cohorts of students from first grade through third grade, with approximately 2,400 students in each grade, were assessed over two school years. Special education students accounted for 10%, and English language learners accounted for 32% of the sample. Each of the schools met specific criteria for student poverty level (at least 69% of students qualifying for free or reduced lunch). Students in the program were given DIBELS measures in fall, winter, and spring, followed by a comprehensive reading test at the end of the year. Participants in first and second grades had two years of data, while study participants in kindergarten and third grade had only one year of data. As part of the Reading First program, each of the students in the study received at least 90 minutes of research-based large group reading instruction and 30 minutes of small group reading instruction per day (Baker et al.). Baker et al.'s study included all subtests of the SAT-10. First grade students were administered the Word Study Skills, Word Reading, Sentence Reading, and Reading Comprehension subtest, and second grade students were administered the Word Study Skills, Reading Vocabulary, and Reading Comprehension subtests of the SAT-10

(Baker et al.). Participants were included if they had at least one ORF and at least one high stakes assessment data point. Correlation coefficients between ORF and first grade SAT-10 scores were .72 in the winter and .82 in the spring. Correlation coefficients for ORF and second grade SAT-10 scores, from the winter of first grade through the spring of second grade, were .63, .72, .72, .79, and .80, respectively.

The study conducted by Baker et al. (2008) had some notable research limitations. The first limitation was that participants were not divided by group membership; as a result, no specific moderating variables could be assessed. In addition, all of the children in the study received interventions. Consequently, it is unknown if the predictive validity varied as a result of intensity and duration of the interventions. Furthermore, the sample only included Reading First schools, which have a majority of low income students. It is unknown if these results would generalize to non-Reading First schools and middle to high income students. Lastly, the study used a longitudinal growth model instead of following one cohort of students over time. This introduces additional sources of error into the study and limits the analysis of predictive validity over time.

Roehrig, Petscher, Nettles, Hudson, and Torgesen (2008) evaluated the relationship of DIBELS ORF probes to the SAT-10 Reading Comprehension component. The study included all 35,207 third grade students attending Florida Reading First schools. Seventy-five percent of the participants qualified for free and reduced lunch, 17% received special education services, and 12% were English language learners. Study participants were grouped into two samples which were controlled for demographic variables. Study participants were given DIBELS ORF probes four times per school year

and the SAT-10 toward the end of the school year. For the first student sample, correlation coefficients between ORF and the SAT-10 comprehension subtest were .68 for the fall, .68 for December, and .71 for February/March when the two measures were given concurrently. For the second sample, correlation coefficients between ORF and the SAT-10 comprehension subtest were .69 for the fall, .68 for December, and .70 for February/March. Scores were not reported for the spring ORF administration (Roehrig et al.).

The sample used in the Roehrig et al.'s (2008) study was similar to Baker et al.'s (2008) study and only included Reading First schools with disproportionate numbers of low income students, which as stated previously, makes generalizability difficult. In addition, Roehrig et al. did not assess differences in performance by subgroup in the SAT-10 analysis. As a result, it is unknown if the predictive validity is stronger for specific study participants. Their analysis only included one year of data for third grade students, which prevents an investigation of whether early intervention can predict performance in later grades. Lastly, the study analysis only included the Reading Comprehension component of the SAT-10. No information was provided on phonemic awareness, phonics, or vocabulary subtests.

Later research on ORF and SAT-10 performance by Wanzek et al. (2010) and Paleologos and Brabham (2011) also focused on low-income students; however, these students were not specifically receiving structured interventions. Wanzek et al.'s analysis of the predictive validity of ORF probes on the SAT-10 included study participants from one school district in Texas made up of six Title 1 elementary schools. Participants

included 461 students in first through third grade. Seventy-four percent of the participants received free or reduced lunch, 14% were limited English proficient, and approximately 5% received special education services. DIBELS ORF measures were administered in the winter and spring for students in first grade. ORF probes were administered in the fall, winter, and spring for students in second and third grade. The Reading Comprehension portion of the SAT-10 was administered at the end the students' third grade year. Correlation coefficients between the ORF probes and the SAT-10 for first grade were .54 for winter and .64 for spring. Correlation coefficients for second grade were .61 for fall, .66 for winter, and .68 for spring. Correlation coefficients for third grade were .68 for fall, .70 for winter, and .69 for spring. All coefficients were large across grade levels (Wanzek et al.).

Resembling Roehrig et al.'s (2008) study, Wanzek et al.'s (2010) study only analyzed the Reading Comprehension portion of the SAT-10, which limits information about other reading components. In addition, the standardized tests were not the actual high stakes tests, but instead were similar tests given by classroom teachers in a mock environment. Children's test performance on the mock tests may differ from their performance in actual testing conditions.

A final study on the predictive validity of ORF and SAT-10 measures was completed by Paleologos and Brabham (2011). Study participants included 215 third grade students from eight public schools in the Southeastern United States. Groups of students were formed based on income status and reading fluency skill level. Students with proficient reading fluency skills included 56 high-income and 56 low-income

students. Students lacking proficient reading skills included 103 low-income students. Students were given the DIBELS ORF and SAT-10 test at the end of their third grade year. Correlation coefficients for proficient readers were .60 for high income students, and .23 for low income students. Correlations coefficients for non-proficient readers were .65; however, when reading vocabulary was controlled for, reading fluency was not a statistically significant predictor of SAT-10 Reading Comprehension scores (Paleologos & Brabham).

The Paleologos and Brabham (2011) study was focused on determining the differences between low income and high income students. Many critical components were excluded as a result. First, only concurrent validity was addressed because the ORF probes and tests were given at the same time. Second, participants were selected based on income status, yielding a non-representative sample because middle class students were not represented and there were no high income poor readers. Lastly, only third grade students were included, thus the results cannot be generalized to other grade levels. Thus generalizability of these findings is limited due to the exclusion of critical components.

Arizona Instrument to Measure Standards

Little is known about the relationship of ORF and the Arizona standardized state test. In a technical report, Wilson (2005) analyzed whether benchmark scores on DIBELS ORF probes could predict a passing score on the AIMS state test. Wilson used a sample of 241 third grade students across three schools, disaggregated by student risk level from the DIBELS standards (Good & Kaminski, 2002). The correlation coefficient

between AIMS and ORF for the overall group was .74. Students were grouped based on their performance compared to benchmark classifications of: at-risk, some-risk, or low-risk. These categories indicate a student needs intensive help, some strategic help, or has met the established benchmark, respectively (Good, Simons, Kame'enui, & Wallin, 2002). When separated by ORF performance, 81.9% of the low-risk group, 51% of the some-risk group, and 7% of the at-risk group were rated as proficient on AIMS. Participant demographics included male (n = 131), female (n = 109), Hispanic (n = 117), white (n = 82), free or reduced lunch (n = 167), pay lunch (n = 70), ELL (n = 65), and non-ELL (n = 175). Students who were in the groups of female, white, not eligible for free/reduced lunch, and not classified as ELL performed better on both ORF probes and AIMS; however, the correlation between ORF and AIMS was relatively consistent across all groups. Prediction estimates per group were .76 for males, .72 for females, .78 for Hispanic students, .68 for white students, .74 for students receiving free lunch, .65 for students who pay for lunch, .78 for ELL students, and .67 for non-ELL students.

One limitation to Wilson's (2005) research is that only third graders from three Reading First schools were included, which reduces the generalizability of the results. Additionally, there is no indication of the time between administrations of the ORF probes and the state assessment. Lastly, specific details were not provided on the measures or the statistical procedures used. Consequently, attempting to validate his findings is difficult, if not impossible.

In an unpublished dissertation, Knight (2007) assessed the use of DIBELS scores in second grade to predict whether students would pass the AIMS DPA (a combination of

criterion items from the AIMS test and norm-referenced items from TerraNova (CTB/McGraw Hill, 2001) at the end of third grade. The reading comprehension items from the TerraNova were taken from the subtests: (a) Basic Understanding, (b) Analyze Text, (c) Evaluate and Extend Meaning, and (d) Identify Reading Strategies. DIBELS scores for 1,450 students across 19 elementary schools were collected at the end of grade 2 and the beginning, middle, and end of grade 3. Study participants were placed into one of three DIBELS benchmark categories, including at-risk, some risk, or low risk. The sample consisted of 68% white, 6% African American, 13.7% Hispanic, 3.1% Native American, and 8.1% Asian students. Study participants consisted of 50.3% female and 49.7% male students. Two percent of these students were ELL, 9.2% were in special education, and 12.2% were identified as gifted.

Analyses were disaggregated by gender and ethnicity. Results of the analyses revealed a moderately large correlation between DIBELS and the reading portion of the AIMS DPA at the end of second grade and third grade, with correlation coefficients of .63 and .62, respectively. This result indicated that ORF scores from second grade and third grade had approximately equivalent predictive validity. Additionally, more white males placed in the DIBELS at-risk category passed AIMS DPA than other groups placed in the at-risk category (Knight, 2007).

Some of the limitations to the study included a relatively homogeneous sample of students and uncertainty as to whether the predictive validity of ORF probes varied as a function of subcomponents included in the AIMS DPA. The study also outlined the intervention component of DIBELS, but did not indicate school participation in these

interventions or if participants differed in performance as a result of reading interventions.

Key Research Limitations

Current research on ORF and high stakes assessments including the SAT-10 and AIMS assessment has some key limitations. First, all the studies used the DIBELS ORF measure, which makes it difficult to determine if these data trends would be observable in other available ORF measures, such as the STEEP. Second, none of the studies grouped participants by intervention status. Study participants in Baker et al. (2008) and Roehrig et al. (2008) studies all received interventions, but there was no comparative sample. No students were listed as receiving interventions in the research conducted by Wanzek et al. (2010), Paleologos and Brabham (2011), Wilson (2005), and Knight (2007). Consequently, there is no information on whether the predictive validity of ORF probes varies as a function of intervention status.

Reasearch Summary on Reading Fluency and High Stakes Tests

Findings across studies assessing the relationship of ORF measures and high stakes tests generally indicated moderate to strong positive correlations between the measures (Baker et al., 2008; Good & Jefferson, 1998; Knight, 2007; Reschly et al., 2009; Roehrig et al., 2008; Wanzek et al., 2010; Wayman et al., 2007; Wilson, 2005; Yeo, 2010) which means that higher scores on ORF measures were related to higher scores on reading achievement tests. The research suggests that ORF was highly correlated with reading comprehension, often above the correlations of direct reading comprehension assessments (Reschly et al.). Correlation coefficients between oral

reading measures and national tests tended to vary more than state tests. The relationship between ORF and national tests also tended to be stronger than the correlations between ORF and state tests (Reschly et al.; Wayman et al.; Yeo). Generally, longer intervals between ORF measures and standardized tests yielded weaker correlation coefficients. The time of year had mixed impact on ORF prediction strength. The spring CBM administration typically had the strongest correlation with high stakes test scores, which were also administered in the spring (Baker et al., 2008; Roehrig et al.; Wanzek et al.).

Language acquisition status had a significant impact on prediction results. Prediction estimates were strong for both ELL and non-ELL students (Reschly et al., 2009; Wilson, 2005; Yeo, 2010). Although students classified as ELL typically had moderate to strong coefficients for reading aloud measures and high stakes test scores, ELL status generally served as a moderator between the curriculum measurement and high stakes test score. For example, students not classified as ELL generally performed better on both ORF probes and high stakes tests.

Study Purpose

Information on the effectiveness of using ORF measures to predict achievement on high stakes assessments is critical for data-based decision making; however, research on the application of ORF predictive validity findings across specific academic abilities and student intervention status is currently limited. The purpose of this study is to address the predictive validity of ORF on both state and norm-referenced tests.

Longitudinal data will be used to assess how soon educators can determine student achievement on high stakes tests. It will also assess if one time of year is superior

to another for making this prediction. This analysis information could be used to target interventions at specific reading deficits. This detailed analysis will also help determine if reading fluency is associated with reading comprehension as some prior research suggests (Fuchs, Fuchs, & Maxwell, 1988). This study will also consider the effect of interventions on the relationship between ORFs and high stakes tests, and specifically address the difference between computer-based versus instructor-led based interventions. Additionally, this research will help determine if curriculum based measurements such as the DIBELS generalize to other ORF measures, specifically the STEEP. Lastly, this study will assess what affect intervention status has upon the predictive abilities of ORF measures.

Research Questions

The research questions for this study include:

1. What is the relationship between ORF probes and the reading composite skill areas assessed by a state reading measure including: (a) Reading Process, (b) Elements of Literature, (c) Comprehending Informational Text, and (d) AIMS DPA Total Reading Score? Additionally, what is the relationship between ORF probes and the reading subtests skill areas assessed by a state reading measure including: (a) Print concepts, (b) Phonics, (c) Vocabulary, (d) Comprehension Strategies, (e) Expository Text, (f) Functional Text, and (g) Persuasive Text?
2. How does time of year relate to student success on high stakes tests?
3. What is the difference between students exposed to computer-based vs. instructor-led intervention and their performance on high stakes tests?

4. What is the predictive validity of ORF probes on a comprehensive state assessment for students disaggregated by intervention status?
5. What is the predictive validity of ORF probes on a norm-referenced measure for students disaggregated by intervention status?

Chapter 2

Method

Participants

Participants included first grade students from four Title 1 elementary schools, located in a large suburban district in Southwest United States. For the 2009 - 2010 school year, approximately 455 first grade students were enrolled in the four selected schools. According to the NCLB act, Title I funding is provided to schools where at least 35% of students come from low-income families. These four schools were selected based on their participation in the district's RTI process. Participant selection criteria included receipt of at least one progress monitoring probe, as well as completion of the AIMS DPA along with the SAT-10 reading component. All students who met these criteria were included in the analysis. From the selected first grade school populations, 143 students were eliminated since they did not meet study criteria. The remaining 312 students made up the participant sample.

The sample included children who were in first grade for the 2009-2010 school year (see Table 1). The sample was relatively equally divided between male and female participants. The ethnicities of study participants based on state racial/ethnic classification categories were primarily White (56%) and Hispanic (28.5%), with the other categories making up 15% of the total sample including Black/African American (5.8%), American Indian/Alaskan Native (4.8%), Asian (2.9%), Native Hawaiian/Other Pacific Islander (.3%), and two or more races (1.6%). Any student identified as a child

with a disability under IDEIA was classified as a special education student. From the sample, 14.4% of students were classified as receiving special education services.

Table 1

Study Participant Demographic Variables

Demographic Variable	*n*	Percent
Gender		
Female	154	49.4
Male	158	50.6
Ethnicity		
White	175	56.1
Hispanic	89	28.5
Black/African American	18	5.8
American Indian/Alaskan Native	15	4.8
Asian	9	2.9
Native Hawaiian/Other Pacific Islander	1	.3
Two or more races	5	1.6
Language acquisition		
English Language Learner	10	3.2
English proficient	302	96.8
Special Education		
Identified Special Education	45	14.4
Not identified Special Education	267	85.6
Discipline referrals		
Prior referrals	38	12.2
No referrals	274	87.8
AIMS DPA proficiency		
Falls Far Below the Standard	10	3.2
Approaches the Standard	61	19.6
Meets the Standard	184	59.0
Exceeds the Standard	57	18.3
Intervention status		
Received interventions	66	21.2
No interventions	246	78.8

Note. AIMS DPA = Arizona Instrument to Measure Standards Dual Purpose Assessment.

For ELL classification, students identified though a home survey as second language learners take the Arizona English Language Learner Assessment (AZELLA; Arizona Department of Education & Harcourt Assessments, 2007). Students who score

below the proficient level are classified as ELL students (Arizona Department of Education, 2011). For the sample, 3.2% were classified as ELL students. Of the participants, 12.2% were recorded as having received discipline referrals. On the AIMS DPA, 3.2% were classified as *Falls Far Below the Standard*, 19.6% as *Approaches the Standard*, 59% as *Meets the Standard*, and 18.3% as *Exceeds the Standard*.

Interventions

Study intervention information was provided by district school psychologists and district RTI documents. Of the study participants, 17.6% of students received Tier II interventions and 10.6% received Tier III interventions across the three years. Overall, 21.2% of the children received some type of intervention from 2009-2012. Participants were placed in a variety of instructor-led and computer-based interventions in Tier II and Tier III.

According to the Scottsdale Unified School District's unpublished *Response to Intervention Manual* (Scottsdale Unified School District, 2013), the district RTI policy indicates universal screenings of all students, followed by focused assessments for students below the 16th percentile, including the "can't do/won't do" assessment. Following these assessments the school RTI teams review the data and determine student placement in interventions. Each school selects interventions targeted at specific academic deficits for Tier II interventions. The movement from the Tier II to Tier III interventions indicates increasing frequency, duration, or instructor qualification. A Tier II intervention is often 30 minutes, 4 days per week, whereas a Tier III intervention is often 45 minutes, 5 days per week. School interventionists often use aspects of the special

education curriculum at this level for students who did not respond to Tier II. All students receive interventions in Tier I, approximately 16% to 20% of students receive Tier II support, and 3% to 5% receive Tier III support. Within the Scottsdale school system, Tier III support is not indicative of special education support services (Scottsdale Unified School District).

The following evidenced-based interventions were utilized in the participants' schools: Fundations (Wilson Language Training, 2006), SuccessMaker (Pearson Education, 2013), Read Naturally (Read Naturally, 2013a), Volunteer support, Mimiosprout Early Reading (Mimio, 2013), Essential Skills (Essential Skills Software, 2012), Scholastic System 44 (Scholastic, 2013), Wilson Reading System (Wilson Language Training, 2004), and Earobics (Earobics, 2007) for Tier II and Wilson Reading System, Fundations (Wilson Language Training, 2006), and SOAR to Success (Houghton Mifflin Harcourt, 2011) for Tier III.

Some study participants received more than one intervention; therefore study numbers are not exclusive. Wilson Reading and Fundations were used in both Tier II and Tier III interventions. All participants received Tier II support before transitioning to Tier III interventions. Students in these Tier III interventions receive more intensive support. The SOAR to Success was used exclusively in Tier III interventions. All three Tier III interventions are components taken from the district's special education curriculum; however, students receiving Tier III interventions were not necessarily receiving special education services (see Table 2).

Table 2

Types of Tier II and Tier III Interventions Given to Study Participants from First Grade through Third Grade

Intervention	*n*	Percent
Tier II Intervention		
Fundations	25	32.5
SuccessMaker	12	15.6
Read Naturally	11	14.3
Volunteer support	10	12.9
Headsprouts	8	10.4
Essential Skills	3	3.9
Scholastic System 44	3	3.9
Wilson Reading System	3	3.9
Earobics	2	2.6
Tier III Interventions		
Wilson Reading System	43	82.6
Fundations	8	15.4
SOAR to Success	1	2.0

Wilson Reading System. The Wilson Reading System is a supplemental reading instruction program for children in second grade or above first published in 1988 (Wilson Language Training, 2004). This system uses direct teaching methods paired with student practice. The Wilson program is designed to be sequential, with students first mastering the designated skills in reading and spelling before moving to the following step. The instruction starts with emphasis on phonological awareness, then progresses to multisyllabic work and finishes with higher level word structure. Each lesson includes work in vocabulary development and comprehension strategies. Initial research on program effectiveness generally showed positive results. O'Connor and Wilson (1995) analyzed the effectiveness of the Wilson program on 220 participants with learning disabilities in grade 3 through grade 12. Students were exposed across the year to an

average of 62 lessons. Students' scores on the Woodcock Reading Mastery Test-Revised (Woodcock, 1987) indicated significant gains across the year with an average of 4.6 grade levels in Word Attack, 1.6 grade levels in Passage Comprehension, and 1.9 grade levels in Total Reading standard scores. An updated study completed by Wood (2002) mirrored positive findings. Results comparing 374 students' pre and post test standard scores on the Woodcock Reading Mastery Test-Revised/Normative Update (Woodcock, 1998) showed student improvement in Word Attack, $t(348) = -22.56$, $p < .001$; Word Identification, $t(350) = -12.35$, $p < .001$; Passage Comprehension, $t(348) = -13.72$, $p < .001$; and the Total Reading Cluster, $t(348) = -15.69$, $p < .001$ following instruction with the Wilson reading system (Wood, 2002).

Fundations. Fundations, a subprogram of the Wilson Reading System, was published in 2002 for children in kindergarten through third grade (Wilson Language Training, 2006). The program was designed to be used in whole classroom instruction, as a targeted intervention for struggling readers, or as a curriculum for students with learning disabilities. The program is based on systematic multi-sensory detailed instruction with integrated practice. Instruction is targeted at phonemic awareness, letter recognition, phonics, and decoding. The program also includes work in vocabulary, fluency, and comprehension strategies. There is no research specifically addressing Fundations, although research on the overall Wilson Reading System was generally positive (O'Connor & Wilson, 1995; Wood, 2002).

Mimiosprout Early Reading. The Mimiosprout Early Reading program (Mimio, 2013) is an online teaching tool developed for children from four to seven years of age.

The program includes 80 lessons with both a computer component and an interactive story component. The program aligns with the National Reading Panel's (2000) five basic skills of phonemic awareness, phonics, fluency, comprehension, and vocabulary. Students must master a level before proceeding to the next. There is no current research on the Mimiosprout Early Reading program, but research on the earlier edition named the Headsprout Early Reading program (Headsprout, 2003) was generally positive (Layng et al.,2003).

SuccessMaker. SuccessMaker is a computer-based intervention program that was developed to address students' academic needs on an individual level (Pearson Education, 2013). The software integrates real-time analysis to continuously customize student learning based on actual performance, and incorporates social studies, science, and interdisciplinary themes. The SuccessMaker program offers specialized instruction when learners encounter difficulties, as well as step-by-step tutorials and scaffolded feedback (Pearson Education). Efficacy research for SuccessMaker is generally positive. According to the Gatti Evaluation (2011), third, fifth and seventh grade students who used the SuccessMaker program outperformed a control group on the Group Reading Assessment Evaluation (GRADE) , a vocabulary and reading comprehension test, and the AIMSWeb Reading CBM and demonstrated more positive attitudes toward reading (Gatti Evaluation, 2011).

Read Naturally. The Read Naturally Program, initiated in 1991, is based on three research-proven strategies for improving reading proficiency; teacher modeling, repeated reading, and progress monitoring (Read Naturally, 2013a). The program addresses

phonemic awareness, phonics, fluency, vocabulary, and comprehension. Specific program steps include (a) student selection of a story , (b) student review of key words, (c) student prediction of the story's subject based on title, key words, and pictures, (d) student identification of difficult words within one minute, (e) graphing of words read correctly within one minute, (f) reading along to the story played on an audio track, (g) timed reading of the story without audio, (h) taking a quiz about the story, (i) assessing the student's capability to read the story within timing and accuracy parameters, (j) graphing scores based on stories that have been passed, (k) retelling the story in writing or verbally, and (l) read a story's key word list (Read Naturally). Efficacy research was significantly positive. A study conducted by Tucker and Jones (2010) compared two groups of students; those who received Read Naturally instruction 30 minutes a day for four days per week, as well as general instruction, and those who only received general instruction. Student accuracy, rate, and fluency were measured using the Gray Oral Reading Test- Fourth Edition (GORT-4; Wiederholt & Bryant, 2001). The results were all statistically significant. The effect size for accuracy was moderate (d = 0.68), and the effect sizes for rate (d = 1.12) and fluency (d = 1.03) were large. (Read Naturally, 2013b).

Essential Skills. Essential Skills is a computer-based intervention program that provides a customized learning experience for each student through the use of auditory, visual, and tactile activities (Essential Skills Software, 2012). The Essential Skills program generates individual lessons based on each student's areas of deficiency. Essential Skills offers three categories of skill development, which are broken down into sub-categories. The main categories include Basic Skills Series, Read to Succeed Series,

and Assessment Series. Efficacy research results for Essential Skills are generally positive, although research is limited (Essential Skills Software).

System 44. System 44 is a program that incorporates a whole-group introduction, followed by small-group rotations, which include instructional computer-based software, instructor-led instruction, and modeled and independent reading (Scholastic, 2013). This program also utilizes various, strategically placed, learning stations within one classroom, which correspond to the whole-group introduction and small-group rotations. System 44 is intended for daily use in classrooms of 10-12 students. The program's efficacy has been thoroughly researched in several states with consistently positive results (Scholastic, 2011).

Earobics. Earobics is a reading intervention program that is comprised of student resources and interactive computer-based software, as well as teacher guides to enhance the professional development of educators (Earobics, 2007). Student resources include books focusing on rhymes, sound starters, sound blends, and leveled readers. In addition, multimedia materials including music, audio cassettes, videotapes, and talking CD-ROMS are included. Lastly, students are provided manipulatives, letter sets, picture/word cards, and letter/sound cards. The program is designed to correlate to all major core reading programs and individual state curriculum standards. Earobics efficacy research is limited; however, a study conducted by Pokorni, Worthington, and Jamison (2004) found that the program was associated with overall gains in phonemic awareness, with significant gains in the area of segmenting phonemes.

SOAR to Success. The SOAR to Success program, originally coined Project Success, uses systematic, scaffolded, small-group instruction (Houghton Mifflin Harcourt, 2011). The SOAR program is based on teaching comprehension strategies for students in grades 3 through 8. The program also targets the basic reading components of: phonemic awareness, phonics, vocabulary, and fluency. Research on the SOAR to Success program is limited; however, a study conducted by Jairam & Kiewra (2010) demonstrated that the using the principles of SOAR in computer based instruction could help enhance student's reading comprehension skills.

Volunteer support. Volunteer support was used to give individual attention to specific students. These interventions were overseen by a school interventionist. Volunteers worked one-on-one with students helping facilitate accurate reading from school books leveled according to skill level. The interventions varied somewhat based on student need. No specific information is available on the types of intervention materials used, intervention methods, or the school volunteers.

Assessment Instruments

System to Enhance Educational Performance. The ORF probes used in this study are one component of the STEEP system. The STEEP system, originally called Problem Validation Screening, and later Screening to Enhance Equitable Placement, was developed to help improve educational services to children and decrease over-identification of special education students (VanDerHeyden, Witt, & Naquin, 2003). Research indicates that STEEP is three times more accurate at identifying struggling students than teacher report (VanDerHeyden et al.). Use of STEEP as it was intended,

has also reduced the number of special education referrals and improved student achievement (VanDerHeyden, Witt, & Gilbertson, 2007). Furthermore, the STEEP system has reduced unnecessary referrals in schools where there are a disproportionate number of high achieving students (VanDerHeyden & Witt, 2005).

The STEEP system includes both math and reading STEEP measures. ORF measures are one minute long individually administered probes. Reading STEEP probe construction was conducted in conjunction with published ORF probe guidelines by Shapiro (1996). The probes were developed in three stages. In the first stage, high and medium frequency words were selected from a database containing 5 million words taken from thousands of children's books. The probes for each grade level were then written using these database words. In the second stage probe readability was established. The Spache readability formula (Spache, 1953) was used for first through third grade, and the Dale-Chall readability formula (Dale & Chall, 1948) was used for fourth grade and up to ensure appropriate level of readability for each probe. In the third stage, probes were evaluated to ensure predictability of student reading skills. Studies examining the psychometric adequacy of the STEEP indicate that the assessment meets or exceeds accepted standards for reliability and validity. Test-retest coefficients for 207 students in grades 1 through 5 ranged from .91 to .95, with a median of .93. Alternate form coefficients ranged from .83 to .88, with a median of .86. Split-half reliability coefficients were .87 for first grade, .71 for second grade, .91 for third grade, .83 for fourth grade, and .87 for fifth grade (Witt, 2007).

The benchmarks used for interpreting the STEEP represent instructional standards for each grade level. Based on the benchmarks cutoff recommendations, ORF scores are used to place students at three levels of proficiency in reading including mastery, instructional, and frustrational. These levels indicate whether a student has no difficulty in reading, may need reading assistance, or has difficulty reading, respectively. Children who score in the frustrational level are given a second screening, referred to as *a Can't do/ Won't do* assessment, with an incentive provided to parse out motivation and skill factors. In the assessment, Children are administered the same ORF measure used in the screening, and told by the examiner if they can outperform their prior score then they will be offered an incentive (Witt, & VanDerHeyeden, 2007). Table 3 illustrates the oral reading fluency end of the year benchmark scores for elementary students, along with corresponding proficiency measures. Information is not made available on beginning or midyear performance criteria.

Table 3

Oral Reading Fluency End of the Year Benchmarks for Students in Elementary School

	Grade Level		
Proficiency Level	First	Second	Third
Frustrational	0-39	0-39	0-69
Instructional	40-60	40-60	70-100
Mastery	61+	61+	101+

Stanford Achievement Test-10th Edition. The SAT-10 is a multiple choice norm-referenced test for students in kindergarten through twelfth grade. The SAT was first published in 1926, and tests in the series have been updated over the years. The

SAT-10 was standardized in 2007 based on a nationally representative sample of students. The SAT-10 scores can be converted into scaled scores, percentile rank scores, stanine scores, grade-equivalent scores, and normal curve equivalent scores (Harcourt Educational Measurement, 2003b).

The SAT includes assessment of skills in Reading, Spelling, Language, Mathematics, Science, and Social Science. A complete or abbreviated battery can be given. An abbreviated battery was given in this particular study. The reading portion measures essential reading skills, including phonemic awareness, decoding, phonics, vocabulary, and comprehension. The specific reading subtests on the complete battery include: Sounds and Letters, Word Study Skills, Word Reading, Reading Vocabulary, Sentence Reading and Reading Comprehension. The subtests that are administered vary based on grade level. The Reading section of the SAT-10 has an alpha reliability rating of .87 (Harcourt Educational Measurement, 2003b).

Arizona's Instrument to Measure Standards. The AIMS (Arizona Department of Education, 2012a) assesses educational content in writing, reading, mathematics, and science for Arizona students. The high school editions of the reading, writing, and mathematics tests, as well as the science test for all grade levels, only include criterion-referenced questions. The AIMS DPA in reading, writing, and mathematics includes both criterion-referenced and norm-referenced questions for elementary students. Students in grades 3, 4, and 8 receive a reading and language AIMS DPA; whereas, students in grades 5, 6, and 7 receive a reading and writing AIMS DPA. The

mathematics AIMS DPA is consistent across grades 3 through 8 (Arizona Department of Education, 2012b).

Items from the AIMS DPA reading test can contribute to a student's score on the criterion-referenced test (CRT), the norm-referenced test (NRT), or both sections. The AIMS CRT for grades 3 and 4 consists of 39 items developed by Arizona teachers, along with 15 norm-referenced SAT-10 reading items that map to the Arizona content standards. The AIMS NRT consists of the 15 norm-referenced CRT questions along with 10 additional SAT-10 reading items (Arizona Department of Education, 2012b).

According to the Arizona Department of Education (2012b), the SAT-10 reading component questions were selected by Pearson's research scientist and psychometricians to match the full form SAT-10 test blueprint and statistical criteria. The 25 SAT-10 reading component questions come from the Reading Comprehension subtest and include the following percentages: (a) 24% Critical Analysis, (b) 24% Initial Understanding, (c) 36% Interpretation, and (d) 16% Strategies. These are similar content percentages to the SAT-10 full form and abbreviated battery, which have 54 questions and 30 questions respectively. SAT-10 reading component results are reported as scale scores, national percentile rank, national stanine, and normal curve equivalents. For 2012 third grade reading scaled scores ranged from 446 to 741. Norms are reported based on the 2007 SAT-10 norms. The SAT-10 reading component, using the 2007 norms, was used in this study (Pearson Education, 2007).

Criterion referenced test scores form scales and are reported across Arizona content standards of reading including: (a) The reading process of print concepts,

phonics, vocabulary, and comprehension strategies; (b) comprehending the elements of literature, as well as the historical and cultural aspects of literary text; and (c) comprehending expository, functional, and persuasive informational text. The strands form the 54 questions from the AIMS DPA third grade test with the following percentage representation from each concept: (a) 7% print concepts, (b) 9% phonics, (c) 11% vocabulary, (d) 17% comprehension strategies, (e) 22% elements of literature, (f) 11% expository text, (g) 11.1% functional text, and (h) 11.1% persuasive text. CRT scores are reported by performance level, pass or fail status, raw score, scaled score, reading concept raw score, and reading concept percent correct score (Arizona Department of Education, 2012b).

The AIMS test items were developed using a multi-segment development process including specifications, initial selection, editing, review, analysis, and selection. All grade levels and content areas were calibrated and scaled with samples typically consisting of the entire Arizona student population. Item Response Theory (IRT) was used for calibrating test items. Item-pattern scoring produces a scale score, which takes into account how many items were answered correctly, the specific items that were answered correctly, and the characteristics of those items. The AIMS scale scores in reading for elementary school students ranges from 200 to 800 depending on the grade level. The AIMS also classifies students into one of four performance categories: Falls Far Below the Standard, Approaches the Standard, Meets the Standard, and Exceeds the Standard. Cut scores are established every year per grade level to determine these groupings. In 2012, third grade reading scores ranged from 200 to 640 with the

following cutoffs: Approaches the Standard standard score (SS) ≥ 379, Meets the Standard SS ≥ 431, and Exceeds the Standard SS ≥ 517 (Arizona Department of Education, 2012b).

The AIMS has strong reliability and validity evidence. Participant data was disaggregated by student ethnicity, gender, ELL status, special education status, socioeconomic status, and migrant status. Internal consistency estimates for the CRT of AIMS ranged from .82 to .93 for grade level students, with ELL receiving the lowest scores across grade levels. Internal consistency estimates for the NRT were lower and ranged from .59 to .85; again, with ELL receiving the lowest scores. Strand and concept internal consistency scores ranged from .43 to .84. Differential Item Functioning (DIF) indicated that there was no significant subgroup bias in the elementary school reading AIMS test. For elementary school, correlation coefficients between the reading CRT and NRT ranged from .69 to .90. Correlation coefficients between reading CRT and language NRT ranged from .76 to .78. Correlation coefficients between reading NRT and language NRT ranged from .70 to .74 (Arizona Department of Education, 2012b).

Procedures

Archival data was collected from two online databases. The first database, AZRTI, includes information from the target school district on RTI; specifically, the STEEP ORF scores from the four selected schools. One cohort of first grade students for the 2009 to 2010 school year was selected to assess the relationship between ORF probes and the selected assessment results over subsequent years. Data was gathered on the ORF scores for study participants' first, second, and third grade school years.

Participants received ORF screenings in the fall, winter, and spring. The fall screenings were completed in August/ September approximately two to four weeks after the start of school. Winter screenings were completed in December/ January in the middle of the school year. Finally, spring screenings were completed in April/ May approximately two to four weeks before the end of school. ORF screening measures were given by school personnel trained in ORF administration.

The second database, Datacentral, is an internal district database which includes demographic information, AIMS DPA scores, and SAT-10 reading component scores for the district students. Information was gathered on the SAT-10 reading component and AIMS DPA performance at the end of the students' third grade year, along with relevant demographic variables. The AIMS DPA and SAT-10 tests are both administered across the month of April. The study was approved as exempt by the Arizona State University (ASU) Institutional Review Board (IRB; see Appendix).

Chapter 3

Results

Data Procedures

Power analysis. Cohen (1988) outlines four essential parameters for statistical analysis including: Statistical power ($1 - \beta$), significance criterion (α), sample size (n), and effect size. Three of the four are typically known and are used to determine the fourth (Cohen). In order to determine study sample size, a power analysis was completed for each research question. Each statistical test will keep an individual alpha level of .05 because each research question was picked based on meaningful interpretations.

Hierarchical multiple regression analyses were used to determine the predictive validity of ORF probes on high stakes tests. The linear multiple regression statistic was analyzed to determine appropriate sample size. The analysis used a fixed model to assess R^2 increase. According to Cohen (1992), a general rule of thumb for the behavioral sciences states that small, medium, and large effect sizes for regression analysis be .02, .15, and .35, respectively. The statistical power level was selected to be .80, and significance level .05 based upon the recommended 4:1 ratio of α to β (Cohen).

Results of the linear multiple regression fixed model power analysis showed that a relatively small effect size (f^2 = .02) would require 791participants. A moderate effect size (f^2 = .15) would require 114 participants, and a large effect size (f^2 = .35) would require 54 participants. A moderate effect size (f^2 = .15) was used for this study because most of the effect sizes found in the CBM prediction literature range from moderate to large. When the significance level is set at .05, power is set at .80, and effect size is set at

15, 114 children were needed to produce the appropriate sample size. Power analysis results using the point biserial correlation model indicated that a small effect size (r = 10 would require 614 participants. For a moderate effect size (r = .30), 64 participants would be needed. For a large effect size (r = .50), 21 participants would be needed. A moderate effect size was selected based on typical effect sizes for ORF and high stakes tests correlations. When the significance level is set at .05, the power level at.80, and the effect size is set at .30, 64 children were needed to achieve the appropriate sample size.

Missing data. In order to be included in the analysis, participants were required to have at least one progress monitoring probe, as well as scores on both the SAT-10 reading component and AIMS DPA assessments. Missing data on the ORF probes was corrected using the regression formula, linear trend at point, which replaces missing values with the predicted value for that point. Participants who had missing data were retained to ensure an accurate data sample, because deleting non-random participants can cause distribution skewness (Tabachnick & Fidell, 1996).

Sample Characteristics

Means and standard deviations were calculated for all demographic variables. Table 4 includes the means and standard deviations for grade 1, grade 2, and grade 3 ORF scores, by demographic variables. On average, females read more words per minute across grade levels than males. ORF scores were similar across ethnicity. Asian and Native Hawaiian/Other Pacific Islander scores were the highest across all three years, whereas Hispanic student scores were lowest across grades 2 and 3. Students with two or more races were lowest in grade 1.

Table 4

Means and Standard Deviations of Oral Reading Fluency Scores

	Grade 1 ORF		Grade 2 ORF		Grade 3 ORF	
Demographic Variable	*M*	SD	M	SD	M	SD
Gender						
Female	57.05	21.03	103.34	28.87	122.13	34.05
Male	50.61	21.32	97.45	33.26	118.45	35.75
Ethnicity						
White	57.49	22.49	108.11	32.93	126.43	38.59
Hispanic	47.33	19.81	86.22	26.63	106.73	28.42
Black/African American	50.28	12.54	91.56	22.25	116.72	18.52
American Indian/Alaskan Native	49.60	14.72	94.07	23.29	122.93	27.20
Asian	60.11	25.64	115.33	19.81	139.44	22.74
Native Hawaiian/Other Pacific Islander[a]	91.00		135.00		137.00	
Two or more races	45.60	5.60	97.20	5.17	112.20	18.58
Language acquisition						
English Language Learner	53.79	21.39	100.36	31.26	120.27	34.92
English proficient	54.52	21.23	101.49	30.76	121.16	34.80
Special Education						
Identified Special Education	49.89	19.15	87.27	34.15	101.67	47.59
Not identified Special Education	54.45	21.71	102.57	30.26	123.40	31.34
Discipline referrals						
Prior referrals	47.29	14.79	88.53	29.93	111.50	31.51
No referrals	54.69	22.02	102.00	31.14	121.48	35.24
AIMS DPA proficiency						
Falls Far Below the Standard	40.10	20.18	71.20	28.72	59.30	38.66
Approaches the Standard	41.87	14.43	76.74	22.41	92.26	26.80
Meets the Standard	52.63	17.62	102.39	27.89	124.09	28.53
Exceeds the Standard	72.70	26.08	124.21	29.38	148.58	26.79
Intervention status						
Received interventions	40.67	13.29	77.15	21.62	87.64	28.40
No interventions	57.31	21.80	106.59	30.54	129.02	31.14

Note. ORF = Oral Reading Fluency words per minute; AIMS DPA = Arizona Instrument to Measure Standards Dual Purpose Assessment.
[a]Standard Deviation not applicable since single participant

English language learners had similar scores when compared to English proficient students. Students identified in any area of special education had lower scores across the three grades than those students not identified as special education. Students with prior discipline referrals also had lower ORF scores across grades. Mean scores consistently

increased across grades with each level of AIMS proficiency from Falls Far Below the Standard, Approaches the Standard, Meets the Standard, and Exceeds the Standard. Expectedly, Students who received interventions had lower ORF scores across grades, compared to those who did not receive interventions.

Table 5 includes the means and standard deviations of SAT-10 reading component scores and AIMS DPA reading scaled score, by demographic variables. Student performance was similar across demographic variables for the SAT-10 reading component scores and AIMS DPA reading scaled score with their results on the ORF scores presented in Table 4. Female participants had higher SAT-10 reading component scores and AIMS DPA reading scaled scores than males. SAT-10 reading component scores and AIMS DPA reading scaled scores were similar across ethnicities. Asian students scored the highest on the SAT-10 reading component score, whereas Native Hawaiian/Other Pacific Islander students scored the highest on the AIMS DPA reading scaled score. Hispanic students scored lowest across both tests. ELL and English proficient students scored similarly across tests. Students identified as special education scored lower across both tests than those not identified as special education. Students with discipline referrals scored lower across both tests than those with no discipline referrals. Students scored higher as they progressed through the AIMS DPA proficiency levels from Falls Far Below the Standard, Approaches the Standard, Meets the Standard, and Exceeds the Standard respectively. Students who received interventions scored lower on both tests than those who did not receive interventions.

Table 5

Means and Standard Deviations of High Stakes Test Scores According to Demographic Variables

	SAT-10 SS		AIMS-DPA SS	
Demographic Variable	M	SD	M	SD
Gender				
Female	637.44	39.51	474.23	47.27
Male	632.56	46.65	464.84	50.12
Ethnicity				
White	647.15	42.99	482.07	47.69
Hispanic	612.31	38.36	444.79	45.13
Black/African American	635.06	36.87	461.72	42.97
American Indian/Alaskan Native	616.80	37.34	457.53	3914
Asian	653.00	22.57	504.44	36.44
Native Hawaiian/Other Pacific Islander[a]	617.00		509.00	
Two or more races	637.00	33.96	461.20	40.54
Language acquisition				
English Language Learner	634.96	43.27	469.48	48.88
English proficient	636.74	42.66	471.67	47.90
Special Education				
Identified Special Education	605.62	42.95	429.91	48.35
Not identified Special Education	639.91	41.40	476.15	45.80
Discipline referrals				
Prior referrals	618.00	44.48	443.66	48.45
No referrals	637.32	42.65	473.06	47.94
AIMS DPA proficiency				
Falls Far Below the Standard	548.10	11.06	362.80	9.91
Approaches the Standard	586.30	19.90	408.66	13.12
Meets the Standard	639.48	27.50	474.02	22.56
Exceeds the Standard	687.70	25.62	538.63	20.23
Intervention status				
Received interventions	605.11	32.80	433.61	37.69
No interventions	642.98	42.27	479.10	47.08

Note. AIMS DPA = Arizona Instrument to Measure Standards Dual Purpose Assessment; SS = Scale Score; SAT-10 = Stanford Achievement Test-10th Edition.
[a]Standard Deviation not applicable since single participant

First Research Question

The first research question was: What is the relationship between ORF probes and the reading composite skill areas assessed by a state reading measure including: (a)

Reading Process, (b) Elements of Literature, (c) Comprehending Informational Text, and (d) AIMS DPA Total Reading Score? Additionally, what is the relationship between ORF probes and the reading subtests skill areas assessed by a state reading measure including: (a) Print concepts, (b) Phonics, (c) Vocabulary, (d) Comprehension Strategies, (e) Expository Text, (f) Functional Text, and (g) Persuasive Text?

Pearson product-moment correlations were calculated to determine the relationships between ORF by grade level and standardized scores on the specific reading skills measured by the AIMS DPA. A mean value was calculated for first, second, and third grade from the ORF values collected across the years to prepare the data for analysis. The grade means were correlated with the AIMS DPA composite scores of reading process, comprehending informational text, and the AIMS DPA total reading score. The ORF scores for the three years were also correlated with the individual AIMS DPA reading score components including: (a) Print concepts, (b) phonics, (c) vocabulary, (d) comprehension strategies, (e) elements of literature, (f) expository text, (g) functional text, and (h) persuasive text. The AIMS DPA component raw scores were converted to z scores with a mean of 0 and a standard deviation of 1.0 in order to accurately compare scores across skill areas. The probability of generating a statistically significant test result increases as the number of tests increases (type 1 error) so the alpha level for each correlation analysis was set at .002 (.05 ÷ 33 analyses) to maintain the familywise error rate at .05.

The results of the correlation analysis between ORF by grade level and measures of reading are presented in Table 6. Mean score analysis indicates that students

performed similarly across skill areas. Participants' scores were highest in the area of phonics and lowest in the area of functional text. Significant correlations were found for each reading component across all grade levels ($p < .001$). Each of the correlation coefficients was positive and greater than or equal to .25. Higher ORF scores were associated with higher reading scores on the components of the AIMS DPA reading test.

Table 6

Means, Standard Deviations, and Correlations between Arizona Instrument to Measure Standards Dual Purpose Assessment Reading Scores and Mean Scores of Oral Reading Fluency by Grade Level

			ORF Year		
AIMS DPA Score Component	M	SD	Grade 1	Grade 2	Grade 3
Reading Process	17.93	4.38	.39*	.49*	.63*
Print concepts	2.83	1.06	.25*	.26*	.33*
Phonics	3.95	1.06	.26*	.30*	.42*
Vocabulary	4.62	1.47	.34*	.45*	.57*
Comprehension Strategies	6.51	1.99	.35*	.45*	.57*
Elements of Literature	7.89	2.71	.45*	.50*	.59*
Comprehending Informational Text	12.06	4.18	.46*	.49*	.58*
Expository Text	4.67	1.38	.34*	.41*	.51*
Functional Text	3.53	1.67	.40*	.42*	.45*
Persuasive Text	3.86	1.87	.41*	.43*	.51*
AIMS DPA Total Reading Score	469.48	48.88	.49*	.54*	.64*

Note. AIMS DPA = Arizona Instrument to Measure Standards Dual Purpose Assessment; ORF = Oral Reading Fluency.
* $p < .001$.

In general, as students progressed from first through third grade, the coefficient magnitudes increased. Correlation coefficients across grades were usually strongly related with the composite scores, including reading process, comprehending informational text, and the AIMS DPA total reading score. Coefficients for the components of reading process were stronger for vocabulary and comprehension strategies, than for the basic reading skills of print concepts and phonics. Medium to

large coefficients were present in elements of literature across grades. Expository, functional, and persuasive texts had similar coefficients although persuasive texts generally yielded the largest coefficients across grades.

Second Research Question

The second research question was: How does time of year relate to student success on high stakes tests, correlation coefficients were computed between ORF administration times and the reading scale scores for AIMS DPA and SAT-10 reading components. A mean value was calculated for fall, winter, and spring from the values collected across the year. The alpha level for each individual test was set at .002 (.05 ÷ 24 analyses) to maintain the familywise error rate at .05.

Table 7 lists the means, standard deviations, and correlations. Mean score analysis indicates an increase in ORF scores across grade years. However, a decrease in ORF scores occurred over the summers between grades, with the summer following the second grade year being the largest drop. The study participants, as a whole, met the ORF STEEP end of the year proficiency standard of mastery. For first and second grade these mastery proficiency level scores are at least 61 words per minute, and for third grade these standards are at least 101 words per minute according to STEEP benchmarks. Significant correlations were found between all times of year and both high stakes test scale scores ($p < .001$). The correlations were all medium to large, represented by coefficients at or above .34. An increase in ORF scores was associated with an increase in AIMS DPA and SAT-10 reading component scores. The ORF scores had a stronger relationship with the AIMS DPA scale scores than with the SAT-10 reading component

scale scores. The largest correlation coefficients were for second grade fall, third grade fall, and third grade spring. Spring administration times were the strongest coefficients for all grades mean scores.

Table 7

Correlations between Oral Reading Fluency Probe Administration Time and High Stakes Test Scores

			Measure	
ORF Administration Time	M	SD	AIMS DPA SS	SAT-10 SS
First Grade Fall	24.76	23.26	.37*	.34*
First Grade Winter	57.55	25.74	.46*	.41*
First Grade Spring	79.14	23.64	.46*	.40*
Second Grade Fall	75.02	31.04	.54*	.53*
Second Grade Winter	104.72	35.70	.50*	.47*
Second Grade Spring	121.37	34.69	.45*	.43*
Third Grade Fall	103.04	34.78	.59*	.57*
Third Grade Winter	122.06	36.98	.51*	.49*
Third Grade Spring	135.71	42.16	.67*	.62*
All Grades Fall Mean	67.63	25.25	.61*	.58*
All Grades Winter Mean	94.77	27.70	.58*	.55*
All Grades Spring Mean	112.09	28.44	.64*	.59*

Note. ORF = Oral Reading Fluency; AIMS DPA = Arizona Instrument to Measure Standards Dual Purpose Assessment; SS = Scale Score; SAT-10 = Stanford Achievement Test-10th Edition.
* $p < .001$.

Third Research Question

The third research question was: What is the difference between students exposed to computer-based versus instructor-led intervention and their performance on high stakes tests? This question addressed the relationship between Tier II intervention types and performance on high stakes tests. Due to limited power, instead of analyzing each type of Tier II intervention participants were exposed to, individual interventions were grouped together in either instructor-led or computer-based interventions. Pearson product-moment correlations were calculated to determine the relationship between the

type of intervention, including computer-based versus instructor-led, and student performance on the AIMS DPA reading score and SAT-10 reading component score (see Table 8).

Table 8

Correlations between Oral Reading Fluency and High Stakes Test Scores by Intervention Type

ORF score	M	SD	AIMS DPA	SAT-10
Computer Based Intervention				
First Grade Fall	9.13	7.77	.27	.23
First Grade Winter	34.72	19.62	.33	.36
First Grade Spring	55.05	17.56	.49	.45
Second Grade Fall	38.61	17.16	.51	.50
Second Grade Winter	67.55	23.97	.51	.48
Second Grade Spring	86.41	29.71	.66*	.63*
Third Grade Fall	75.52	28.67	.81*	.80*
Third Grade Winter	99.54	25.57	.18	.17
Third Grade Spring	106.74	33.76	.65*	.62
First Grade Mean	32.91	13.11	.44	.43
Second Grade Mean	64.13	20.26	.68*	.65*
Third Grade Mean	93.96	25.03	.66*	.64*
Instructor-led Intervention				
First Grade Fall	13.04	9.27	.18	.08
First Grade Winter	51.80	13.24	.43	.34
First Grade Spring	72.87	15.27	.44	.34
Second Grade Fall	53.80	23.21	.38	.26
Second Grade Winter	98.40	18.57	.43	.34
Second Grade Spring	113.30	22.65	.43	.34
Third Grade Fall	70.96	33.67	.59*	.51
Third Grade Winter	86.52	32.01	.43	.39
Third Grade Spring	90.70	38.08	.70*	.59*
First Grade Mean	45.87	10.67	.44	.33
Second Grade Mean	88.43	17.61	.50	.37
Third Grade Mean	82.72	30.80	.65*	.57*

Note. ORF = Oral Reading Fluency words per minute; AIMS DPA = Arizona Instrument to Measure Standards Dual Purpose Assessment; SAT-10 = Stanford Achievement Test-10th Edition.
* $p < .001$

Computer-based interventions included: (a) Mimiosprout Early Reading program, (b) SuccessMaker, (c) Essential Skills, (d) System 44, and (e) Earobics. Instructor-led interventions included (a) Wilson Reading System, (b) Fundations, (c) Read Naturally

Program, and (d) volunteer support. For the first grade, 17% of participants received computer-based interventions and 83% received instructor-led interventions. In the second grade, 43% received computer-based interventions and 57% received instructor-led interventions. In the third grade, 35% received computer-based interventions and 65% received instructor-led interventions. The alpha level for each individual test was set at .001 (.05 ÷ 48 analyses) to maintain the familywise error rate at .05.

ORF mean score comparisons indicated that ORF mean scores were higher for students in the instructor-led interventions for first and second grade, but were higher for the participants in the computer-based interventions for third grade. Statistically significance was found in 14 of the correlations. Each significant correlation was large and greater than or equal to .57 ($p < .001$). The correlations for first grade and the beginning of second grade tended to be lower and not significant. Correlations for computer-based interventions tended to be stronger than those for instructor-led interventions. Correlations for students in the computer-based intervention group were large and statistically significant approaching the end of second grade compared to the start of third grade for the instructor-led students.

Fourth Research Question

The fourth research question was: What is the predictive validity of ORF probes on a comprehensive state assessment for students disaggregated by intervention status? This analysis was disaggregated by students who received interventions and those who did not receive interventions.

AIMS DPA no interventions. Preliminary analyses were conducted for the non-intervention group to ensure there were no violations of the assumptions of linearity and homoscedasticity. A scatterplot of the standardized residuals and predictive variables was completed (see Figure 1). The residuals represent the difference between the actual and predicted values based on the regression equation. In general, the scatterplot points were randomly distributed across the Y axis and X axis. They also were generally linear and consistently spread. This indicated appropriate linearity and homoscedasticity.

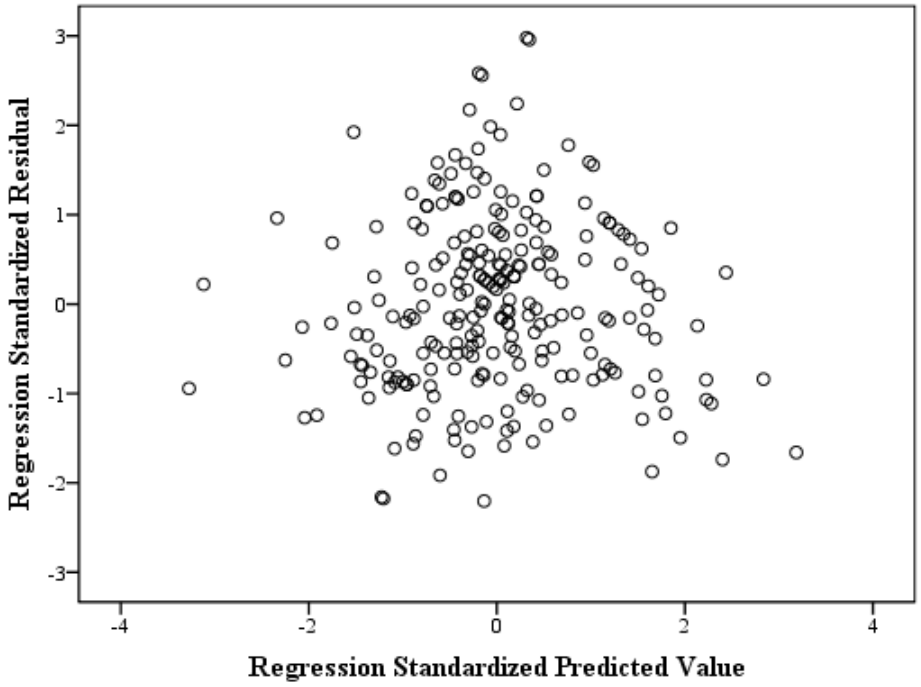

Figure 1. Scatterplot depicting the association between the standardized residuals and the standardized predictive values from the hierarchical regression analysis predicting the Arizona Instrument to Measure Standards Dual Purpose Assessment reading score from oral reading fluency scores for students who did not receive interventions.

Hierarchical multiple regression analyses were used to assess the predictive validity of ORF scores on reading achievement. Independent variables included the ORF

curriculum-based measures from the fall, winter, and spring administrations across the three years. The criterion variables included the AIMS DPA and SAT-10 reading component scores. Each additional block was evaluated after controlling for the impact of the previously considered blocks. The models were disaggregated by intervention status.

The first group of analyses focused on predictive validity for AIMS DPA scores. ORF probe scores were added from the fall of first grade through the spring of third grade in a forward fashion to determine their effectiveness in predicting AIMS DPA achievement. This was repeated for students who were receiving interventions. The second group of analyses focused on the predictive validity of ORF probes on SAT-10 reading component achievement scores. ORF probe information was added from the fall of first grade through the spring of third grade in a forward fashion. This was repeated for all children receiving interventions.

The Variance-inflation factor (VIF) was referenced to test for multicollinearity. The VIF represents how much the variance changes for regression coefficients in the scenario where predictors are uncorrelated (Cohen, Cohen, West, & Aiken, 2003). The higher a VIF score, the higher the correlation is with other predictors. VIFs of 10 or more are generally considered too large to interpret analysis results (Cohen et al.). The VIF tests for multicollinearity indicated that a low level of multicollinearity was present (VIF = 2.50, 3.57, 1.89, 3.20, 4.24, 3.63, 2.42, 2.86, and 3.61 from the fall of first grade through the spring of third grade, respectively).

Means, standard deviations, and intercorrelations are listed in Table 9. Mean score analysis indicates an increase in ORF scores across grades. Correlations were significant across all ORF predictors and the AIMS-DPA variable ($p < .05$). All correlations were positive and greater than or equal to .32. Correlations tended to be strongest when the two measures were close in time, for example, third grade winter and spring administrations.

Table 9

Means, Standard Deviations, and Intercorrelations between Arizona Instrument to Measure Standards Dual Purpose Assessment Reading Scores with Oral Reading Fluency Measures for Students who did not Receive Interventions

Variable	M	SD	1	2	3	4	5	6	7	8	9
AIMS Predictor	479.10	47.08	.32*	.46*	.42*	.49*	.49*	.39*	.50*	.41*	.59*
1. F Grade 1	28.07	24.79		.77*	.49*	.58*	.58*	.53*	.39*	.39*	.41*
2. W Grade 1	60.77	26.60			.68*	.72*	.72*	.71*	.52*	.51*	.56*
3. S Grade 1	83.17	23.09				.60*	.58*	.56*	.50*	.36*	.44*
4. F Grade 2	82.79	28.35					.79*	.74*	.64*	.56*	.64*
5. W Grade 2	110.22	35.95						.83*	.56*	.60*	.65*
6. S Grade 2	126.72	34.47							.61*	.66*	.67*
7. F Grade 3	111.04	31.69								.60*	.71*
8. W Grade 3	130.59	34.07									.78*
9. S Grade 3	145.47	38.56									

Note. AIMS = Arizona Instrument to Measure Standards Dual Purpose Assessment; F = Fall: W = Winter; S = Spring.
* $p < .05$.

The results of the hierarchical regression analysis predicting the AIMS DPA reading score from ORF across three years for students who did not receive interventions are reported in Table 10. The alpha level for each individual comparison was set at .006 (.05 ÷ 9 analyses) to maintain the familywise error rate at .05. ORF probe scores were added from the fall of first grade through the spring of third grade in a forward fashion to determine their effectiveness in predicting AIMS DPA achievement.

Table 10

Hierarchical Multiple Regression Analysis Predicting Arizona Instrument to Measure Standards Dual Purpose Assessment Reading Scores with Oral Reading Fluency Measures for Students who did not Receive Interventions

Step and predictor variable	*R*	R^2	ΔR^2	ΔF	*B*	*SE*	B	*T*
Grade 1	.48	.23	.23	24.31**				
Fall					-.11	.17	-.06	-.67
Winter					.65	.19	.37**	3.45
Spring					.41	.16	.20^{+}	2.57
Grade 2	.54	.30	.06	7.23**				
Fall					.35	.16	.21^{+}	2.16
Winter					.41	.15	.31*	2.77
Spring					-.25	.14	-.18	-1.75
Grade 3	.65	.43	.13	18.28**				
Fall					.19	.11	.13	1.66
Winter					-.18	.12	-.13	-1.54
Spring 2					.63	.11	.51**	5.48

Note. Betas reported are those from the step at which the variable was entered into the equation; SE = Standard Error.
$^{+}$ $p < .05$. * $p < .006$. ** $p < .001$.

The three predictors from first grade testing times were all entered in the first step of the hierarchical multiple regression analysis. This model was statistically significant $F(3, 242) = 24.31$; $p < .001$ and explained 23% of variance in AIMS DPA reading scores ($R^2 = .23$). The second step entered an additional three predictors from the second grade ORF testing times. This model was also statistically significant $\Delta F(3, 239) = 7.23$; $p < .001$. The introduction of second grade ORF scores explained an additional 6% variance in AIMS DPA scores, after controlling for first year ORF scores ($\Delta R^2 = .06$). In the final model, the addition of third grade ORF scores explained an additional 13% of variance over the first two years of predictors ($\Delta R^2 = .13$), which was statistically significant, $\Delta F(3, 236) = 18.28$; $p < .001$. After entry of all ORF scores, the total variance explained by the model was 43% ($R^2 = .43$). Three of the predictors were statistically significant

including first grade winter (β = .37, p = .001), second grade winter (β = .31, p = .006), and third grade spring (β = .51, p < .001). A change in third grade spring ORF scores had a stronger effect upon the AIMS DPA scores than the other coefficients.

AIMS DPA received interventions. A residual scatterplot was analyzed to assess model assumptions (see Figure 2).

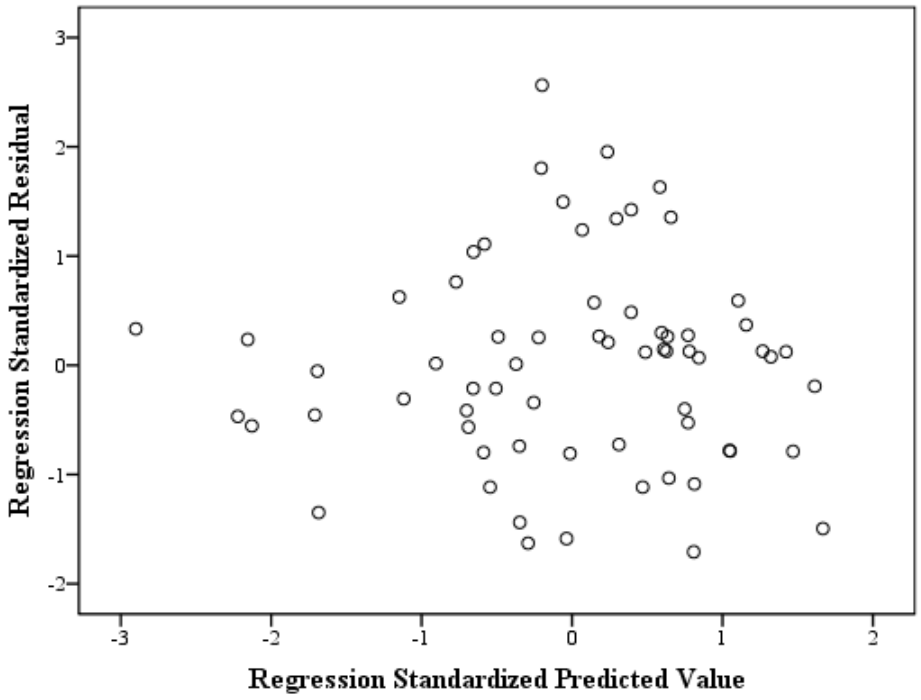

Figure 2. Scatterplot depicting the association between the standardized residuals and the standardized predictive values from the hierarchical regression analysis predicting the Arizona Instrument to Measure Standards Dual Purpose Assessment reading score from oral reading fluency scores for students who received interventions.

In general, the points were randomly distributed across the Y axis and X axis, linear, and consistently spread which indicated the model had appropriate linearity and homoscedasticity. The VIF tests for multicollinearity indicated that a low level of

multicollinearity was present (VIF = 1.31, 2.33, 2.48, 1.43, 4.61, 4.44, 3.35, 3.86, and 4.00, respectively).

Means, standard deviations, and intercorrelations are listed in Table 11. Generally ORF scores increased over grade levels. Correlations tended to be statistically significant, strong, and positive when two measures were close together in time ($p < .05$). Statistically significant negative correlations were present between third grade winter ORF scores and most of the first grade and second grade ORF scores. The results of the hierarchical regression predicting the AIMS DPA reading score from ORF across three years for students who received interventions are reported in Table 12. The alpha level for each individual comparison was set at .006 (.05 ÷ 9 analyses) to maintain the familywise error rate at .05.

Table 11

Means, Standard Deviations, and Intercorrelations between Arizona Instrument to Measure Standards Dual Purpose Assessment Reading Scores with Oral Reading Fluency Measures for Students who Received Interventions

Variable	M	SD	1	2	3	4	5	6	7	8	9
AIMS	433.61	37.69	.08	.08	.20	.20	.12	.31*	.58*	.44*	.61*
Predictor											
1. F Grade 1	12.42	8.77		.42*	.48*	.44*	.43*	.42*	.18	-.23*	-.02
2. W Grade 1	45.54	17.78			.75*	.32*	.72*	.59*	-.01	-.36*	-.17
3. S Grade 1	64.11	19.34				.46*	.82*	.83*	.15	-.25*	-.04
4. F Grade 2	46.06	22.35					.45*	.43*	.33*	.07	.18
5. W Grade 2	84.24	26.16						.83*	.07	-.30*	-.14
6. S Grade 2	101.41	27.66							.20	-.19	-.00
7. F Grade 3	73.24	29.33								.67*	.78*
8. W Grade 3	90.24	29.38									.80*
9. S Grade 3	99.33	34.57									

Note. AIMS = Arizona Instrument to Measure Standards Dual Purpose Assessment; F = Fall: W = Winter; S = Spring.
* $p < .05$.

The first model that included the three predictors from first grade testing was not statistically significant $F(3, 62) = 1.08$; $p = .364$. The second step entered an additional three predictors from the second grade ORF probes. This model was statistically significant $\Delta F(3, 59) = 3.01$; $p = .037$. The introduction of second grade ORF scores explained an additional 13% variance in AIMS DPA scores, after controlling for first year ORF scores ($\Delta R^2 = .13$). The model including third grade ORF scores was statistically significant, $\Delta F(3, 56) = 12.02$; $p < .001$ and explained an additional 32% of variance over the first two years of predictors. All years of ORF scores explained 50% of the variance ($R^2 = .50$). None of the specific predictors were statistically significant at the $p < .006$ level.

Table 12

Hierarchical Multiple Regression Analysis Predicting Arizona Instrument to Measure Standards Dual Purpose Assessment Reading Scores with Oral Reading Fluency Measures for Students who Received Interventions

Step and predictor variable	*R*	R^2	ΔR^2	ΔF	*B*	*SE*	B	*T*
Grade 1	.22	.05	.05	1.08				
Fall					-.05	.61	-.01	-.09
Winter					-.33	.40	-.016	-.84
Spring					.63	.38	.32	1.65
Grade 2	.42	.18	.13	3.01^{+}				
Fall					.27	.24	.16	1.13
Winter					-.70	.37	-.49	-1.92
Spring					.94	.34	.69^{+}	2.76
Grade 3	.71	.50	.32	12.02**				
Fall					.23	.22	.18	1.01
Winter					-.06	.24	-.04	-.24
Spring					.55	.21	.50^{+}	2.66

Note. Betas reported are those from the step at which the variable was entered into the equation; SE = Standard Error.
$^{+}$ $p < .05$. * $p < .001$. ** $p < .001$.

Fifth Research Question

The fifth research question is: What is the predictive validity of ORF probes on a norm-referenced measure for students disaggregated by intervention status?

SAT-10 no interventions. The residual analysis indicated that generally points were randomly distributed across the Y axis and X axis, linear, and consistently spread, which indicated appropriate linearity and homoscedasticity (see Figure 3). The VIF tests for multicollinearity indicated that a low level of multicollinearity was present (VIF = 2.50, 3.57, 1.89, 3.20, 4.24, 3.63, 2.42, 2.86, and 3.61 from the first grade fall through the spring of third grade, respectively).

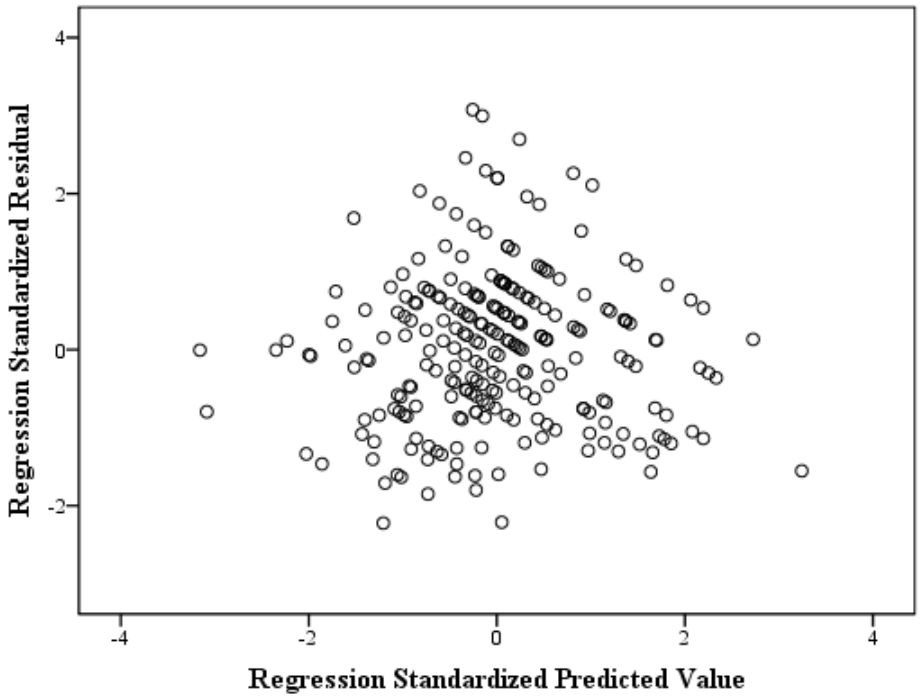

Figure 3. Scatterplot depicting the association between the standardized residuals and the standardized predictive values from the hierarchical regression analysis predicting the Stanford Achievement Test-10th Edition reading component score from oral reading fluency scores for students who did not receive interventions.

Means, standard deviations, and intercorrelations are listed in Table 13. Mean score analysis indicates an increase in ORF scores across grades. Correlations were significant across all ORF predictors and the SAT-10 reading component variable ($p <$ .05). All correlations were positive and greater than or equal to .30. Correlations tended to be strongest when the two measures were close in time.

The results of the hierarchical regression analysis predicting the SAT-10 reading component score from ORF scores across three years for students who did not receive interventions are reported in Table 14. The alpha level for each individual comparison was set at .006 (.05 ÷ 9 analyses) to maintain the familywise error rate at .05. The three predictors from first grade testing were entered in the first step of hierarchical multiple regression analysis. This model was statistically significant $F(3, 242) = 17.55$; $p < .001$ and explained 18% of variance in SAT-10 reading component scores ($R^2 = .18$).

Table 13

Means, Standard Deviations, and Intercorrelations between Stanford Achievement Test-10th Edition Reading Component Scores with Oral Reading Fluency Measures for Students who did not Receive Interventions

Variable	M	SD	1	2	3	4	5	6	7	8	9
AIMS	642.98	42.27	.30*	.41*	.36*	.49*	.47*	.39*	.47*	.39*	.56*
Predictor											
1. F Grade 1	28.07	24.79		.77*	.49*	.58*	.58*	.53*	.39*	.39*	.41*
2. W Grade 1	60.77	26.60			.68*	.72*	.72*	.71*	.52*	.51*	.56*
3. S Grade 1	83.17	23.09				.60*	.58*	.56*	.50*	.36*	.44*
4. F Grade 2	82.79	28.35					.79*	.74*	.64*	.56*	.64*
5. W Grade 2	110.22	35.95						.83*	.56*	.60*	.65*
6. S Grade 2	126.72	34.47							.61*	.66*	.67*
7. F Grade 3	111.04	31.69								.60*	.71*
8. W Grade 3	130.59	34.07									.78*
9. S Grade 3	145.47	38.56									

Note. AIMS = Arizona Instrument to Measure Standards Dual Purpose Assessment; F = Fall: W = Winter; S = Spring.
* $p < .05$.

The second step entered an additional three predictors from second grade ORF. This model was statistically significant $\Delta F(3, 239) = 9.65$; $p < .001$ and explained an additional 9% variance in SAT-10 reading component scores ($\Delta R^2 = .09$). The addition of third grade ORF scores was statistically significant $\Delta F(3, 236) = 13.53$; $p < .001$ and explained an additional 11% of variance over the first two years of predictors ($\Delta R^2 = .11$). The total variance explained by the model was 38% ($R^2 = .38$). Three of the predictors were statistically significant including first grade winter ($\beta = .33$, $p = .003$), second grade fall ($\beta = .30$, $p = .003$), and third grade spring ($\beta = .46$, $p < .001$). Results indicated that a change in third grade spring ORF scores had a stronger effect on the AIMS DPA scores than the other coefficients.

Table 14

Hierarchical Multiple Regression Analysis Predicting Stanford Achievement Test-10th Edition Reading Component Scores with Oral Reading Fluency Measures for Students who did not Receive Interventions

Step and predictor variable	*R*	R^2	ΔR^2	ΔF	*B*	*SE*	β	*t*
Grade 1	.42	.18	.18	17.55**				
Fall					-.04	.16	-.025	-.27
Winter					.52	.18	.33*	2.98
Spring					.27	.15	.15	1.83
Grade 2	.52	.27	.09	9.65**				
Fall					.45	.15	.30*	3.04
Winter					.33	.13	.28^{+}	2.44
Spring					-.17	.13	-.14	-1.30
Grade 3	.61	.38	.11	13.53**				
Fall					.17	.11	.12	1.52
Winter					-.15	.11	-.12	-1.34
Spring					.50	.11	.46**	4.66

Note. Betas reported are those from the step at which the variable was entered into the equation; SE = Standard Error.
$^{+}$ $p < .05$. * $p < .006$. ** $p < .001$.

SAT-10 received interventions. Points on the residual scatterplot for the analysis predicting the SAT-10 reading component score from ORF scores were randomly distributed across the Y axis and X axis, linear, and consistently spread which indicated appropriate linearity and homoscedasticity (see Figure 4). The VIF tests for multicollinearity indicated that a low level of multicollinearity was present (VIF = 1.31, 2.33, 2.48, 1.43, 4.61, 4.44, 3.35, 3.86, and 4.00, respectively).

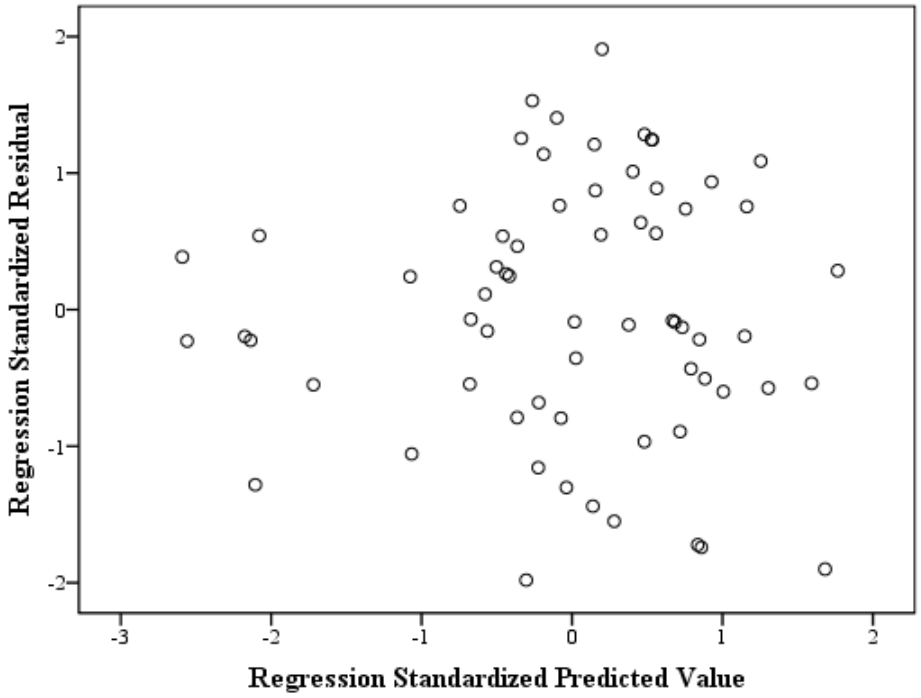

Figure 4. Scatterplot depicting the association between the standardized residuals and the standardized predictive values from the hierarchical regression analysis predicting the Stanford Achievement Test-10th Edition reading component score from oral reading fluency scores for students who received interventions.

Means, standard deviations, and intercorrelations are listed in Table 15. Generally, ORF scores increased across grade levels. Correlations tended to be statistically significant, strong, and positive when two measures were close together in

time ($p < .05$). Statistically significant negative correlations were present between winter 2011 and most of the first grade and second grade ORF scores.

The results of the hierarchical regression predicting the SAT-10 reading component score from ORF across three years for students who received interventions are reported in Table 16.

Table 15

Means, Standard Deviations, and Intercorrelations between Stanford Achievement Test-10th Edition Reading Component Scores with Oral Reading Fluency Measures for Students who Received Interventions

Variable	M	SD	1	2	3	4	5	6	7	8	9
AIMS	605.12	32.80	-.01	.01	.10	.07	.01	.19	.55*	.43*	.56*
Predictor											
1. F Grade 1	12.42	8.77		.42*	.48*	.44*	.43*	.42*	.18	-.23*	-.02
2. W Grade 1	45.54	17.78			.75*	.32*	.72*	.59*	-.01	-.36*	-.17
3. S Grade 1	64.11	19.34				.46*	.82*	.83*	.15	-.25*	-.04
4. F Grade 2	46.06	22.35					.45*	.43*	.33*	.07	.18
5. W Grade 2	84.24	26.16						.83*	.07	-.30*	1.14
6. S Grade 2	101.41	27.66							.20	-.20	-.01
7. F Grade 3	73.24	29.33								.67*	.78*
8. W Grade 3	90.24	29.38									.80*
9. S Grade 3	99.33	34.57									

Note. AIMS = Arizona Instrument to Measure Standards Dual Purpose Assessment; F = Fall: W = Winter; S = Spring.
* $p < .05$.

The alpha level for each individual comparison was set at .006 (.05 ÷ 9 analyses) to maintain the familywise error rate at .05. The three predictors from first grade testing were entered in the first step of the hierarchical multiple regression analysis. These three predictors were not statistically significant $F(3, 62) = 1.08$; $p = .698$. The three predictors from second grade testing were entered in the second step of hierarchical multiple regression analysis, which were also not statistically significant $\Delta F(3, 59) = 2.04$; $p = .092$. The addition of third grade ORF scores in the final step was statistically significant

$\Delta F(3, 56) = 9.44$; $p < .001$) and explained an additional 30% of variance over the first two years of predictors ($\Delta R^2 = .30$). All years of ORF scores explained 41% of the variance ($R^2 = .41$). None of the predictors were statistically significant at the $p < .006$ level.

Table 16

Hierarchical Multiple Regression Analysis Predicting Stanford Achievement Test-10th Edition Reading Component Scores with Oral Reading Fluency Measures for Students who Received Interventions

Step and predictor variable	*R*	R^2	ΔR^2	ΔF	*B*	*SE*	B	*t*
Grade 1	.15	.02	.02	.48				
Fall					-.22	.54	-.06	-.40
Winter					-.28	.35	-.15	-.79
Spring					.40	.34	.23	1.20
Grade 2	.34	.11	.09	2.04				
Fall					.12	.22	.07	.50
Winter					-.64	.33	-.51	-1.93
Spring					.70	.31	.59+	2.29
Grade 3	.64	.41	.30	9.44**				
Fall					.32	.21	.28	1.51
Winter					-.08	.23	-.07	-.36
Spring					.38	.20	.40	1.97

Note. Betas reported are those from the step at which the variable was entered into the equation; SE = Standard Error.
+ $p < .05$. * $p < .006$. ** $p < .001$.

Chapter 4

Discussion

Research Summary

The RTI model is used by school teams to help ensure that all students receive appropriate instruction and adequate support. Scores on ORF CBMs are used in the RTI model to determine whether a student needs, or would benefit from, reading interventions. Using ORF scores to determine student success on high stakes testing is beneficial in order to quickly target deficit skills in schools, classrooms, or individual students. Previous research on the effectiveness of using ORF scores to predict future success on high stakes tests yielded moderate to large positive correlations (Good & Jefferson, 1998; Reschly et al., 2009; Wayman et al., 2007). Research on the applications of ORF predictive validity findings across specific academic abilities and student interventions is more limited.

This study was conducted in order to determine the relationship between ORF skills and high stakes testing. Specifically, the study was designed to assess the predictive validity of ORF measures on AIMS DPA and SAT-10 reading component scores, taking into consideration participant intervention status. In addition, the study was designed to gather information on the relationship of ORF to specific aspects of reading, including comprehension and basic decoding skills. The study also aimed to assess the relationship between ORF and intervention type including computer-based versus instructor-led. Lastly, the study was planned to assess the relationship between the time of year and high stakes assessments in reading.

Study participants included 312 students from four elementary schools located in a large suburban district in Southwest United States. Measures in this study included ORF probes from the iSTEEP system. Reading comprehension questions in analysis, understanding, interpretation, and strategies were assessed based on the SAT-10 reading component scale score. The AIMS DPA was used as the state measure in reading, decoding, and comprehension.

Reading Skills

Conclusions. The results of the correlation analysis between ORF by grade level and different measures of reading from the AIMS DPA were all statistically significant and greater than or equal to .25. This mirrors past research, which indicates reading fluency accurately measures comprehension (Fuchs et al., 2001; Reschly et al., 2009; Shinn et al., 1992). Based on coefficients, ORF actually had a stronger relationship to reading comprehension than to basic reading skills. Testing results also indicated that ORF had a medium to large effect in the areas of understanding elements of literature and comprehending informational texts. Since literature and informational texts permeate curriculums other than the language arts, the predictive validity of ORF may be useful across subjects.

Study results support prior research that correlation coefficients were generally stronger with short intervals between the ORF measures and the high stakes tests measures (Baker et al., 2008; Roehrig et al., 2008; Wanzek et al., 2010). This study also supported findings in prior research that suggest that correlation coefficients could vary in strength depending on the area being assessed such as decoding skills and

comprehension skills (Good & Jefferson, 1998; Reschly et al., 2009; Wayman et al., 2007; Yeo, 2010). Correlation coefficients ranged from small, for first grade print concepts ($r = .25$, $p < .001$), to large, for third grade AIMS DPA reading total scores ($r = .64$, $p < .001$).

Limitations and future research. A limitation to the study findings is the use of a single state test to determine the relationship of ORF to reading skills. These findings may not generalize to other state tests or norm-referenced measures. Additionally, because there are only 54 questions in the reading section of the AIMS DPA, the specific strands include a limited number of questions. This lowers the internal reliability of the individual strand scores. Future research should continue to address the relationship between ORF and aspects of reading ability, but should include more complete measures of the reading subcomponents.

Implications. As with past research, ORF continues to be an accurate measure of reading comprehension (Fuchs et al., 2001; Reschly et al., 2009; Shinn et al., 1992). This information encourages the selection of ORF measures even when comprehension skills are the primary interest. Student performance on ORF measures can also be used when trying to assess a student's performance on high stakes reading comprehension tests. Additionally, ORF may not only be useful in screening for the learning disability category of reading fluency, but also for reading comprehension.

This research indicates that ORF is also moderately to strongly related to understanding literary and informational texts, which are integrated across the curriculum. These brief ORF measures may be a feasible way for school teams to obtain

a snapshot of how a student may perform across subjects, including social studies and science. This information can be used in leveling students or targeting classroom support.

Time of Year

Conclusions. The results of the correlation analysis between ORF administration times and the reading scale scores from the high stakes assessments were all statistically significant. Results assessing administration time mean scores for fall, winter, and spring supported prior research that the spring administration time was usually the strongest testing point related to high stakes tests (Baker et al., 2008; Roehrig et al., 2008; Wanzek et al., 2010). However, when analyzed across the years, results were mixed. For example, fall scores were sometimes as highly correlated to the high stakes test scores as the spring scores were. Winter ORF scores generally had the smallest correlation to the AIMS DPA and SAT-10 reading component scale scores.

The AIMS DPA and ORF correlation coefficients from this study were similar to Knight's (2007) results for the end of third grade, but were lower than Knight's correlation coefficients for the end of second grade, with scores of .67 versus .62 and .45 versus .63 for both studies, respectively. Study coefficients ranged from .37 to .67 on the AIMS DPA and from .34 to .62 on the SAT-10 reading component. These coefficients were similar in strength to prior research on high stakes tests (Good & Jefferson, 1998; Reschly et al., 2009; Wayman et al., 2007; Yeo, 2010).

Limitations and future research. A limitation of the time of year analysis was the inability to control for extraneous variables. Other factors that could influence the

relationship of ORF to high stakes tests, such as varying ORF administration times across schools, were not controlled for because the study was based on archival data. Future research should be conducted by holding constant extraneous variables and only varying the ORF probe administration time.

Implications. Spring ORF scores typically had the strongest relationship to high stakes tests, whereas winter ORF scores generally had the weakest. Consequently, the resources of the RTI teams may be best focused primarily on fall and spring screening.

Type of Intervention

Conclusions. An analysis of individual reading interventions could not be conducted due to limited sample size, thus all computer-based interventions were aggregated and compared to non-computer based interventions. The results of the correlation analysis examining the difference between computer-based versus instructor-led interventions had some notable findings. ORF mean scores were consistently higher across first and second grade for students who received instructor-led interventions. However, by third grade, students in the computer-based interventions had higher ORF scores across fall, winter, and spring screenings.

Across intervention types, the correlations for first grade and the beginning of second grade tended to be smaller and non-significant. This indicates that these grades were less predictive of SAT-10 reading component or AIMS DPA reading standard score performance, regardless of the intervention used.

Correlations, in general, were higher for students receiving computer-based interventions. This indicates that the relationship between high stakes tests and ORF is

stronger when students are in computer-based interventions. These stronger correlations could result from the more stringent standardization procedures available through computer-based interventions. Correlations also tended to be significant by the second grade for those in computer-based interventions compared to the third grade for the instructor-led intervention participants. This indicates that a relationship between interventions and high stakes tests can be predicted earlier in students receiving computer-based support.

Limitations and future research. One limitation to the study is that, it cannot be determined if the differences in scores indicate a higher growth rate for students between interventions because a correlation analysis based on archived data was used. Additionally, study participants were exposed to interventions for varying amount of time, making this relationship unclear. Future research should be conducted using experimental manipulations to assess if computer-based interventions lead to increased growth. They could also help determine if one intervention is superior to another based on participant age and grade level. Participant intervention status should be held constant in order to better understand the relationship.

Another limitation is that it is unclear if computer-based interventions are better at predicting future test performance, though there is a stronger relationship between the computer-based interventions and high stakes tests. Future research should include experimental manipulations to determine whether student performance in computer-based interventions accurately predicts high stakes test performance. This research could help

illuminate whether students make more consistent gains in computer-based interventions or in small-group based interventions.

Implications. The preliminary indication that computer-based interventions may lead to increased student ORF skills may warrant the consideration of the use of computer-based interventions rather than instructor-led interventions. This mimics past research which shows computer-based interventions may have advantages over instructor-led only interventions (Saine et al., 2011).

High Stakes Test Prediction

AIMS DPA conclusions. The predictive validity of ORF on the AIMS DPA reading score was generally strong for students who did not receive interventions. The ORF for all three years was able to explain 43% of the overall AIMS DPA score variance. The ORF scores from the first grade could explain 23% of the variance, an additional 6% variance for the second grade, and a further 13% variance for the third grade. In general, the one minute reading probe from the first grade was able to explain a relatively substantial amount of the score fluctuation in the comprehensive reading assessment more than two years after it was given. In fact, the addition of the next two years combined did not explain as much variance as the first year alone.

The results for students who received interventions were dissimilar from those for students who did not receive the interventions. The only ORF administration year that was statistically significant for the intervention sample was third grade when the AIMS DPA and ORF measures were administered within a two month period rather than one or two years away from the measures for second and first grade, respectively. The first

grade ORF scores only accounted for 5% of the variance, compared to 23% of the variance for the students who did not receive interventions. The second year accounted for 13% of the variance, compared to 6% of the variance for the students who did not get the interventions. Third grade alone accounted for 32% of the variance in AIMS DPA scores for students who received interventions, compared to only 13% for students who did not receive interventions. The aggregate intervention sample model accounted for 50% of the variance in AIMS DPA scores, which is greater than the 43% variance accounted for by the prior model. Thus, the use of interventions, especially standardized computer-based interventions may moderate the relationship of ORF scores and high stakes testing performance.

These results do not necessarily align with prior research on the AIMS DPA that had similar coefficients across groups (Wilson, 2005). However, intervention status was not addressed in past research.

SAT-10 conclusions. Results of the SAT-10 reading component followed a similar pattern but were less predictive when compared to the AIMS DPA results. The first grade ORF scores accounted for 18% of the total variance in the SAT-10 reading component score for children who did not receive interventions. The second grade ORF scores explained 9% of variance above and beyond that of the first year. The third grade ORF scores explained 11% of variance when holding the effects of the first two years constant. The model aggregate explained 38% of the total variance in the SAT-10 reading component scores. The additions of the second and third grade ORF scores had almost the same predictive strength as the first grade year alone.

Once again, the first two years were not significant in explaining SAT-10 reading component variance for students who did not receive interventions. The first grade ORF scores accounted for 2% of the variance compared to a 19% score variance from the non-intervention sample. The second grade ORF scores accounted for an additional 9% of the variance, which was the same as the non-interventional model. Third grade alone accounted for 30% of the total variance in the SAT-10 reading component score, which was much higher than the 11% variance accounted for by the non-interventional sample model. This aggregate model accounted for 41% of the SAT-10 reading component score variance. Similar to the AIMS DPA results, predictions based on ORF scores from first grade varied significantly for students who received interventions and those who didn't. For students who did not receive interventions, the SAT-10 scores predicted better cross time when the measures of administration times were farther apart. For the students who received interventions, SAT-10 scores were predicted best in the third grade when administration times of the SAT-10 and ORF measures were closer in time.

Limitations and future research. A limitation of the analysis is the generalizability of these findings to other students. This study included a homogeneous sample from four elementary schools in one school district. As a result, the sample is limited in demographic and regional representation. Future research should be conducted with a more representative sample in order to determine if these results are consistent across other districts and states.

A limitation of the regression analysis was the high stakes measures selected. There was an overlap between the AIMS DPA and SAT-10 reading component tests.

This overlap prevents us from drawing conclusions about a criterion-referenced versus norm-referenced measure. In addition, these results may not correlate as well with state tests that do not include SAT-10 measures in their assessments. Further, only the SAT-10 reading component paired with the AIMS DPA was analyzed. The predictive validity of using ORF probes may be different on the full battery. With the upcoming adoption of the Partnership for Assessment of Readiness for College and Careers (PARCC) Mid-Year, Performance-Based, and End-of-Year Assessments (Partnership for Assessment of Readiness for College and Careers, 2013) the relationship between ORF and the PARCC reading tests should be explored in future research.

Another limitation to the regression analysis was that students were only grouped as either receiving or not receiving interventions. Differences were not assessed based on specific program intervention types because this was an archival analysis as opposed to a controlled experiment. However, it appears that computer-based interventions may be more efficacious, but as previously stated this needs to be explored further in using a controlled experiment, perhaps by comparing types of computer-based interventions to each other as well as to a control group.

In addition, the length of intervention exposure was not controlled for in the analysis. Some of the students received one semester of interventions, whereas others received three years. A larger sample size would enable the analysis of prediction based on the length of intervention.

Implications. Consistent with prior research, ORF is strongly related to students' AIMS and SAT-10 results (Baker et al., 2008; Knight, 2007; Roehrig et al., 2008;

Wanzek et al., 2010; Wilson, 2005). ORF screenings may be useful in understanding which students are likely to pass their third grade high stakes tests due to the strength of this predictive validity. Outside of targeted interventions, this information could help inform administrators of necessary curriculum adjustments across classrooms or across the school. Changes could be made quickly and effectively for students flagged as at risk for failing their third grade high stakes test.

Regarding high stakes tests, student performance is best predicted based on first grade ORF scores for students not in interventions and on third grade ORF scores for those receiving interventions. One reason for these disparate findings could be that students who receive interventions may change over the years and, consequently, their first grade scores are not as representative of how they will perform by the third grade. Hypothetically, if the reading interventions work, the student receiving them will have an improved performance on the high stakes tests they receive in third grade. Students not in interventions may follow a typical trajectory. Therefore, their ORF performance in first grade is representative of their eventual performance on the AIMS test, in comparison to students who do receive interventions who do not appear to follow a typical trajectory. School teams should take this performance into consideration when making high stakes decisions such as retention and special education placement. Teams should collect data from students receiving interventions on a frequent basis; however, they may still not be able to develop an accurate projection of long-term outcomes for these students. Intervention specialists should work collaboratively with teachers and administrators to better understand these students.

Study Summary

This research study investigated ORF scores from 312 study participants across their first through third grade years. Study participants were selected from four elementary schools involved in RTI. The participants ORF screening measures were compared to their performance on high stakes test scores.

The first study area of interest was aimed at determining the association between the ORF probes and the reading skill areas measured by a state achievement test. Correlation coefficients between the ORF probes and the state achievement test measure were stronger between ORF scores and comprehension measures than between ORF scores and basic reading measures. Based on these results, school intervention teams may be aided by selecting frequency measures when comprehension skills are also of interest. Future research should be completed with more complete measures of reading decoding and comprehension skills.

The second area of interests was focused on determining which ORF screening administration time was best at predicting high stakes test scores. The spring ORF screening scores were most strongly associated with high stakes test score performance. These findings suggest a need to focus resources on spring and fall screenings and less on winter screenings, which have the weakest relationship to high stakes test scores. Future research should hold constant ORF administration time to establish experimental control.

The third study area of interest was focused on assessing if computer-based or instructor-led interventions were more closely related to performance on high stakes test scores. Results indicated student who were in computer-based interventions tended to

have stronger correlation coefficients between ORF probe and high stakes test than those receiving instructor-led interventions. These results indicate that computer-based interventions may need to be incorporated in order to ensure outcome measure relationship strength. Future research should analyze specific intervention types to determine the most effective interventions.

The last areas of interest included assessing the predictive validity of ORF probes on high stakes achievement tests disaggregated by intervention status. Study results indicated that students' ORF screening scores accounted for the most model variance in first grade when they did not receive interventions and in third grade when they did receive interventions. These results suggest the need to consider the unique trajectory of students receiving interventions when making high stakes decisions. Future research should be completed to determine if these results generalize to other intervention samples.

References

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

Arizona Department of Education. (2011). *Arizona English Language Learner Assessment (AZELLA) form AZ-2 technical report*. Retrieved from http://www.azed.gov/standards-development-assessment/files/2012/02/azellaformaz-2technicalreport-february2011.pdf

Arizona Department of Education. (2012a). *Arizona's Instrument to Measure Standards.* San Antonio, Texas: Pearson.

Arizona Department of Education. (2012b). *Arizona's Instrument to Measure Standards 2012 technical report*. Retrieved from http://www.azed.gov/standards-development-assessment/files/2012/12/aims_tech_report_2012_final.pdf

Arizona Department of Education & Harcourt Assessments. (2007). *Arizona English Language Learner Assessment.* Phoenix, AZ: Arizona Department of Education & Harcourt Assessments.

Arizona READS. Arizona Revised Statutes§15-704 (2001).

Arizona Reading Success Task Force. (2000). Arizona reading success task force: Report and recommendations. Retrieved from http://azmemory.azlibrary.gov/utils/getfile/collection/statepubs/id/8475/filename/8767.pdf

Arizona LEARNS. Arizona Revised Statutes § 15-241(2001).

Baker, S., Smolkowski, K., Katz, R., Fien, H., Seeley, J., Kame'enui, E., & Beck, T. C. (2008). Reading fluency as a predictor of reading proficiency in low-performing high poverty schools. *School Psychology Review, 37,* 18-37.

Blok, H., Oostdam, R., Otter, M. E., & Overmaat, M. (2002). Computer-assisted instruction in support of beginning reading instruction: A review. *Review of Educational Research, 72,* 101-130.

Brown-Chidsey, R. & Steege, M. W. (2006). Introduction: What is response to intervention (RTI)?. In *Response to intervention: Principle and strategies for effective practice* (pp. 1-12). New York: Guilford.

CTB/McGraw-Hill. (2001). *TerraNova, the second edition, California Achievement Tests.* Monterey, CA: Author.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 122,* 155-159.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. London: Lawrence Erlbaum Associates.

Coulter, G., K., Shavin, K., & Gichuru, M. (2009). Oral reading fluency: Accuracy of assessing errors and classification of readers using a 1-min timed reading sample. *Preventing School Failure, 54,* 71-76.

Dale, E. & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin, 27,* 11-20.

Deno, S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual.* Reston, VA: Council for Exceptional Children.

Earbonics. (2007). *Pre-k to 3rd grade programs*. Retrieved from http://www.earobics.com/solutions/programs.php

Education for All Handicapped Children Act of 1975, 20 U.S.C. § 1400 et seq. (1975).

Elementary and Secondary Education Act of 1965, 20 U.S.C. 6301 et seq. (1965).

Essential Skills Software. (2012). *Programs*. Retrieved from https://essentialskills.net/products

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

Fuchs, L., & Fuchs, D. (1997). Use of curriculum-based measurement in identifying students with disabilities. *Focus on Exceptional Children, 30,* 1-16.

Fuchs, D., Fuchs, L. S., & Compton, D. (2012). SMART RTI: Next generation approach to multilevel prevention. *Exceptional Children, 78,* 263-279.

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

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

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

Gatti Evaluation. (2011). *Pearson Successmaker reading efficacy study.* Retrieved from http://assets.pearsonschool.com/asset_mgr/current/20136/sm-reading-rct-report1.pdf

Good, R., & Jefferson, G. (1998). Contemporary perspectives on curriculum-based measurement validity. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 61−88). New York: Guilford Press.

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

Good, R. H., Simons, D., Kame'enui, E., & Wallin, J. (2002). *Summary of decision rules for intensive, strategic, and benchmark instructional recommendations in kindergarten through third grade* (Technical Report No. 11). Eugene, OR: University of Oregon.

Harcourt Educational Measurement (2003). *Stanford Achievement Test* (10th ed.). San Antonio, TX: Harcourt Assessment.

Harcourt Educational Measurement (2003). *Technical manual: Stanford Achievement Test* (10th ed.). San Antonio, TX: Harcourt Assessment.

Headsprout (2003). Headsprout Early Reading. Retrieved from http://www.headsprout.com/

Herr, S. M. (2003). A brief history of the field of learning disabilities. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbooks of learning disabilities* (pp. 57-73). New York: Guilford.

Hintze, J. M., Christ, T. J., & Methe, S. A. (2006). Curriculum-based assessment. *Psychology in the Schools, 43,* 45-56.

Houghton Mifflin Harcourt. (2011). *Soar to success 1999-2006.* Retrieved from http://www.hmhschool.com/store/ProductCatalogController?cmd=Browse&subcmd=LoadDetail&ID=1000000000000281&sortEntriesBy=SEQ_NAME&sortProductsBy=SEQ_TITLE&division=S01#description

Hudson, R., Lane, H., & Pullen, P. (2005). Reading fluency assessment and instruction: What, why, and how? *The Reading Teacher, 58,* 702-714.

Individuals with Disabilities Education Act of 1990, 20 U.S.C. § 1400 et seq. (1990).

Individuals with Disabilities Education Act Amendments of 1997, 20 U.S.C. §§ 1400 et seq. (1997).

Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C. § 1400 et seq. (2004).

Jairam, D., & Kiewra, K. A. (2010). Helping students soar to success on computers: An investigation on the SOAR study method for computer-based learning. *Journal of Educational Psychology, 102,* 601-614.

Knight, K. G. (2007). *Predicting third grade success on the AIMS Dual Purpose Assessment using second grade oral reading fluency scores* (Doctoral dissertation). Retrieved from http://search.proquest.com.ezproxy1.lib.asu.edu/docview/304895581

Layng, T. V. J., Twyman, J. S., & Stikeleather, G. (2003). Headsprout Early Reading: Reliably teaching children to read. *Behavioral Technology Today, 3,* 7-20.

Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What is it and why do it. In M. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guilford Press.

Mimio. (2013). Mimiosprout Early Reading. Retrieved from http://www.headsprout.com/Pearson Education. (2013). *SuccessMaker: A digital learning curriculum*. Retrieved from http://www.mimio.com/en-NA/Products/MimioSprout-Early-Reading.aspx

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

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

No Child Left Behind Act of 2001, 20 U.S.C. 70 § 6301 et seq. (2002).

O'Connor, J., & Wilson, B. (1995). Effectiveness of the Wilson reading system used in public school training. In Shaywitz, S., & Shaywitz, B. (2005). Dyslexia (specific reading disability). *Biological Psychiatry, 57,* 1301−1309.

Paleologos, T., & Brabham, E. (2011). The effectiveness of DIBELS oral reading fluency for predicting reading comprehension of high- and low-income students. *Reading Psychology, 32,* 54-74.

Partnership for Assessment of Readiness for College and Careers. (2013). *PARCC accessibility features and accommodations manual: Guidance for districts and decision-making teams to ensure that PARCCmid-year, performance-based, and end-of-year assessments produce valid results for all students.* Retrieved from http://www.parcconline.org/parcc-assessment-policies

Pearson Education. (2007). *Stanford Achievement Test Series, Tenth Edition: 2007 Spring Supplemental Multilevel Norms.* San Antonio, TX: Author.

Pearson Education. (2013). *SuccessMaker: A digital learning curriculum*. Retrieved from http://www.mimio.com/en-NA/Products/MimioSprout-Early-Reading.aspx

Pokorni, J. L., Worthington, C. K., & Jamison, P. J. (2004). Phonological awareness intervention: Comparison of Fast for Word, Earobics, and LiPS. *The Journal of Educational Research, 97,* 147-157.

Read Naturally. (2013a). *How it works*. Retrieved from http://www.readnaturally.com/approach/steps.htm

Read Naturally. (2013b). *Read Naturally control group studies*. Retrieved from http://www.readnaturally.com/approach/case3.htm

Regtvoort, A. G. F. M., & van der Leij, A. (2007). Early intervention with children of dyslexic parents: Effects of computer-based reading instruction at home on literacy acquisition. *Learning and Individual Differences, 17,* 35-53.

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. (2009). Curriculum-Based Measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47,* 427-469.

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

Ryder, J. F., Tunmer, W. E., & Greaney, K. T. (2008). Explicit instruction on 73 phonemic awareness and phonetically based decoding skills as an intervention

strategy for struggling readers in whole language classrooms. *Reading and Writing, 21,* 349-369.

Saine, N. L., Lerkkanen, M., Ahonen, T., Tolvanen, A., & Lyytinen, H. (2011). Computer-assisted remedial reading intervention for school beginners at risk for reading disability. *Child Development, 82,* 1013-1028.

Scholastic. (2011). *Raising reading achievement for America's most challenged older students* . Retrieved from http://system44.scholastic.com/pdf/research/System44_Results_FINAL.pdf

Scholastic. (2013). *System 44*. Retrieved from http://system44.scholastic.com/about/instructional-model

Scottsdale Unified School District. (2013). *Response to Intervention manual.* Manuscript in preparation.

Shapiro, E. S. (1996). Academic skills problems: Direct assessment and intervention. *The Reading Teacher, 32,* 403-408.

Shinn, M. T. (1995). Identifying and defining academic problems: CBM screening and eligibility procedures. In M.R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 90-129). New York: Guilford Press.

Shinn, M. R. (2007). Identifying students at risk, monitoring performance, and determining eligibility within response to intervention: Research on education and benefit from academic intervention. *School Psychology Review, 36,* 601-617.

Shinn, M. R. Good, R., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*, 459-479.

Spache, G. (1953). A new readability formula for primary-grade reading materials. *The Elementary School Journal 53*, 410-413.

Stansfield, W. D. (2011). Educational curriculum standards & standardized educational tests: Comparing apples & oranges? *The American Biology Teacher, 73,* 389-394.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42,* 795-819.

Tabachnick, B., & Fidell, L. (1996). *Using multivariate statistics* (3th ed.). New York: Herper Collins College Publishers.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study. *Review of Educational Research, 81,* 4-28.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Herron, J., & Lindamood, P. (2010). Computer-assisted instruction to prevent early reading difficulties in students at risk for dyslexia: Outcomes from two instructional approaches. *Annals of Dyslexia, 60,* 40-56.

Tucker, C., & Jones, D. (2010). Response to intervention: Increasing fluency, rate, and accuracy for students at risk for reading failure. *National Forum of Educational Administration and Supervision Journal*, *28*, 28-47.

U. S. Department of Education. (2006). Assistance to states for the education of children with disabilities and preschool grants for children with disabilities (34 CFR Parts 300 and 301). *Federal Register, 71,* 46540-46845.

VanDerHeyden, A. M., & Witt, J. C. (2005). Quantifying context in assessment: Capturing the effect of base rates on teacher referral and a problem-solving model of identification. *School Psychology Review, 34,* 161-183.

VanDerHeyden, A. M., Witt, J. C., & Gilbertson, D. A. (2007). Multi-year evaluation of the effects of a Response to Intervention (RTI) model on identification of children for special education. *Journal of School Psychology, 45,* 225-256.

VanDerHeyden, A. M., Witt, J. C., & Naquin, G. (2003). The development and validation of a process for screening and referrals to special education. *School Psychology Review, 32,* 204-227.

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

Wanzek, J., Roberts, G., Linan-Thompson, S., Vaughn, S., Woodruff, A. L., & Murray, C. S. (2010). Differences in relationship of oral reading fluency and high-stakes measures of reading comprehension. *Assessment for Effective Intervention, 35,* 67-77.

Wiederholt J. L., & Bryant, B. R. (2001). *Gray Oral Reading Test-IV.* Austin, TX: Pro-Ed.

Wilson, J. (2005). *The relationship of Dynamic Indicators of Basic Early Literacy Skills (DIBELS) oral reading fluency to performance on Arizona Instrument to Measure Standards (AIMS)* (Technical Report). Tempe, AZ: Tempe School District No.3.

Wilson Language Training. (2004). *The Wilson Reading System.* Retrieved from http://www.wilsonlanguage.com.

Wilson Language Training. (2006). *Wilson Fundations.* Retrieved from http://www.wilsonlanguage.com.

Witt, J. (2007). STEEP CBM screening and intervention for at risk children (Data file and code book). Retrieved from http://www.joewitt.org/.

Witt, J. C., & VanDerHeyden, A. M. (2007). The System to Enhance Educational Performance (STEEP): Using science to improve achievement. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of response to intervention* (pp. 343-353). New York: Springer.

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

Wood, F. (2002). *Wilson literacy solutions: Evidence of effectiveness.* Unpublished manuscript. Wake Forest University, Winston-Salem, NC.

Woodcock, R. W. (1987). *Woodcock Reading Mastery Tests–Revised*. Circle Pines, MN: American Guidance Service.

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

Yang, J. (2006). *A meta-analysis of the effects of interventions to increase reading fluency among elementary school students.* (Doctoral dissertation). Retrieved from http://login.ezproxy1.lib.asu.edu/login?url=http://search.proquest.com/docview/304989870?accountid=4485. (304989870).

Yell, M. L., & Dragow, E. (2007). Assessment for eligibility under IDEIA and the 2006 regulations. *Assessment for Effective Intervention, 32,* 202-213.

APPENDIX

IRB APPROVAL LETTER

ASU Knowledge Enterprise Development

Office of Research Integrity and Assurance

To:	Linda Caterino Kulhavy EDB
From:	Mark Roosa, Chair Soc Beh IRB
Date:	04/09/2012
Committee Action:	**Exemption Granted**
IRB Action Date:	04/09/2012
IRB Protocol #:	1203007660
Study Title:	The relationship of curriculum based measurement probes on standardized achievement tests

The above-referenced protocol is considered exempt after review by the Institutional Review Board pursuant to Federal regulations, 45 CFR Part 46.101(b)(1) .

This part of the federal regulations requires that the information be recorded by investigators in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects. It is necessary that the information obtained not be such that if disclosed outside the research, it could reasonably place the subjects at risk of criminal or civil liability, or be damaging to the subjects' financial standing, employability, or reputation.

You should retain a copy of this letter for your records.