UNIVERSITY OF CALIFORNIA
RIVERSIDE

Passage Equivalency and Predictive Validity of Oral Reading Fluency Measures

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

Doctor of Philosophy

in

Education

by

Christopher Jason Checca

March 2012

Dissertation Committee:
Dr. Michael L. Vanderwood, Chairperson
Dr. Rollanda O'Connor
Dr. Gregory Palardy

The Dissertation of Christopher Jason Checca is approved:

_____

_____

_____
Committee Chairperson

University of California, Riverside

Acknowledgements

I wish to thank my family: Robert, Karen and Katie Checca. Each of you has helped me become the person I am today. I thank you all for your support, inspiration, guidance and above all your love.

There have been too many faculty and staff that have been a tremendous influence on my professional and personal development to list here. However, I would like to give special thanks to my advisor, Dr. Vanderwood, for his continued support and guidance over the last six years. I would like to thank Dr. O'Connor has been a tremendous resource of information about reading and empirical research. Finally, I would like to think Dr. Kevin Apple, my advisor at James Madison University, who worked tirelessly to help me transition from undergraduate to graduate school.

Dedication

For my wife, Carrie, who has supported me from the beginning. For my son, Zachary, who is the reason for all that I do. I love you both.

ABSTRACT OF THE DISSERTATION

Passage Equivalency and Predictive Validity of Oral Reading Fluency Measures

by

Christopher Jason Checca

Doctor of Philosophy, Graduate Program in Education
University of California, Riverside, March 2012
Dr. Michael L. Vanderwood, Chairperson

The use of oral reading fluency (ORF) passages within a Response to Intervention (RTI) framework is examined. Significant limitations within the current ORF research are discussed. The passage equivalency and readability scores for DIBELS Next, AIMSweb, and a school district's curriculum's ORF passages are evaluated using Generalizability Theory and readability formulas. Multiple regression is used to analyze the contribution of ORF progress monitoring passages for predicting the California Standards Test (CST). The optimal number of ORF passages to administer is also examined. Participants consisted of third and fifth grade students from an urban school district in Southern California. Results indicate that readability formulas provide wide range of scores for individual passages but rank sets of passages fairly equally. Results also indicate that ORF passages have high levels of reliability and variance attributable to student skill. Finally, results also indicate that the addition of progress monitoring did not increase the

predictive validity of the CSTs. The implications, limitations, and future direction of research are discussed.

Table of Contents

Passage Equivalency and Predictive Validity of Oral Reading Fluency Measures

Response to Intervention (RTI) is a comprehensive educational model that uses proactive and preventative strategies to identify students at risk for academic failure (Gersten et al., 2008). The RTI model is part of a paradigm shift in educational policy and research. The proactive and preventative strategies within RTI include identifying at-risk students through early educational screening, evaluation of instructional match to student needs, and the use of evidence-based instructional and intervention strategies to inform decisions (Fuchs, Mock, Morgan, & Young, 2003; Gersten et al., 2008).

The RTI model most often comprises a 3-tiered system of implementation. At each stage of the 3-tiered model, data are used to evaluate a student's level of performance and, when applicable, growth. Within the 3-tier model, Tier 1 refers to the general education setting that all students initially receive. Students who are identified as at risk for academic failure, typically by universal screening measures, are considered for placement in interventions that provide additional support in their areas of deficiency. A student placed in Tier 2 may receive small group instruction focusing on his or her needs. In a Tier 2 intervention, a goal is chosen based upon the student's current level of performance and expected rate of growth. Progress monitoring data are collected to determine the student's responsiveness to the intervention and assess the progress towards the student's goal. Students who fail to show adequate growth in a Tier 2 intervention are considered for placement in a Tier 3 intervention. Tier 3 interventions are typically more intensive in both frequency and duration and are often conducted 1-to-1 instead of in small groups (Gersten et al., 2008). At every stage of the RTI model, a

student's placement is based upon data collected from universal screening or progress monitoring.

Universal Screening and Progress Monitoring

Universal screening and progress monitoring are two of the key components within the RTI model (Fuchs, Mock, Morgan, & Young, 2003). Universal screening refers to the assessment of all students using research-based measures and is one of the proactive and preventative features of RTI. Unlike traditional educational models that wait for student failure to begin assessing student abilities (Gresham, 2007), the RTI model allows for the early identification of students at-risk for academic failure before they actually experience that failure. In recent years, curriculum based measures (CBM) have been the assessment tool typically used for both of these assessments.

The CBMs used in progress monitoring are similar to those used in universal screening. However, unlike universal screening, progress monitoring is typically only administered to those students who have already been identified as at-risk. For example, if a student is placed in a Tier 2 intervention after being identified as at-risk, progress monitoring assessments are given throughout the intervention in order to determine the student's responsiveness to the intervention. Thus, universal screening and progress monitoring, and specifically the measures used for each of these components, are of critical importance. The data derived from these measures influence the type of educational services a child receives.

The type of measurement administered is dependent upon the student's grade. For example, beginning in the first grade, students are administered oral reading fluency

(ORF) passages as part of the universal screening. The student's scores are then compared to absolute benchmarks that are used as an indicator of the student's progress towards end-of-year goals. A student who is demonstrating ORF scores below the absolute benchmark scores is considered to be at risk for later failure.

One example of a universal screener is Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 1996). DIBELS provides tri-annual universal screenings with benchmark scores that are used to categorize a student's level of performance. At second grade and above, the universal screener that is used by DIBELS is ORF. Based upon a student's ORF score, a student is classified as 'low risk,' 'some risk,' or 'at risk.' These categories are used as indicators of a student's progress and, for an underperforming student, are an early indicator of the need for intervention. Thus, ORF data are critical pieces of information used in the decision making process within an RTI framework.

Efficacy of RTI

There is an extensive body of research that lends support to the efficacy of an early and proactive educational design. Research has demonstrated repeatedly that early identification of and intervention for students at risk of academic failure leads to positive academic outcomes (Vellutino, Scanlon, & Zhang, 2007). Juel (1988) found that 88% of students who were poor readers in first grade remained poor readers in fourth grade. Francis, Shaywitz, Stuebing, Shaywitz and Fletcher (1996) found that 75% of the children who were poor readers in third grade were still poor readers in 9th grade. Francis et al. found that, for these children, a deficit model and not the 'catch up' or lag model

was supported. Children who are good readers tend to read more than those that are poor readers thereby increasing the gap between skilled and non-skilled readers (Rayner, Foorman, Perfetti, Pesetsky, & Seldenberg, 2002; Stanovich, 1986). Speece and Ritchey (2005) found similar discrepancies across a shorter time-span. The authors found that children identified as at-risk in the beginning of first grade were reading significantly poorer than non-identified students by the end of first grade. The at-risk students were reading less than half as many words per minute and at half the rate of growth as their normally developing peers.

Torgesen, Alexander, Wagner, Rashotee, Voeller, and Conway (2001) examined the effectiveness of intensive early intervention for students demonstrating significant reading difficulties. The authors identified 60 children between the ages of eight and ten with severe reading disabilities and assigned them to one of two reading interventions. In one intervention the students focused on articulation cues, auditory and visual cues, and 85% of instruction time focused on phonemic decoding. In the other intervention group the students focused on sight words, phonemic spelling, and 20% of instruction time focused on phonemic decoding. Each intervention group had two phases of implementation. The first phase consisted of 50 minutes of 1-on-1 intervention for eight to nine weeks and the second phase consisted of in-class generalization work using the same intervention techniques. Assessments were conducted after the second phase and then again one and two years later. The intervention results showed significant gains for all students at post test, regardless of intervention grouping. In addition, 50% of the children were in the average level at post test and 40% of the students were returned to

general education from special education services. There was also significant growth in receptive and expressive language as well as increase in phonological memory. The authors state that their results indicate that some students, who are labeled as learning disabled and in need of special education under traditional identification techniques, may be caught up to average reading levels if given an intense reading intervention. Thus, when ineffective instruction is not ruled out, students may be misidentified and placed in an inappropriate educational setting. This is both disadvantageous and potentially harmful to the child as well as a misappropriation of limited educational resources.

The more efficient use of limited resources is one of the potential benefits to the RTI model (Gersten et al., 2008). Burns, Appleton, and Stehouwer (2005) conducted a meta-analysis to synthesize existing research on the effectiveness of large-scale implementations of RTI. The authors found that across 21 articles meeting their criteria, there were strong effect sizes for field-based RTI models (mean effect size [MES] = 1.38), and university-implemented RTI models (MES = 1.04). Similarly, student outcomes (MES = .96) and systemic outcomes (i.e. number of referrals, students identified as LD, time spent in special education, and number of students retained; MES = 1.53) showed significant improvements using an RTI model. Other research supports the RTI model's focus on early identification and interventions as being an effective means to identify students at risk for academic failure (Torgesen, et al, 2001; Vellutino, Scanlon, & Tanzman, 1998).

Vellutino et al. (1998) identified groups of poor and normal reading kindergarten students via teacher recommendation and standardized assessment scores. In first grade

the poor reading students were randomly assigned to receive intervention from tutors or within their home school. The interventions lasted between 1-2 semesters, dependent upon student response, and tutored students were grouped based upon the level of their responsiveness to the intervention (i.e. students that demonstrated Very Good, Good, Limited, and Very Limited growth during intervention were grouped together). The students' reading abilities were assessed through the third grade. The authors found that early and intensive interventions (i.e. students receiving 1-on-1 tutoring) led to a significant reduction (67.1%) of students identified as disabled readers. Thus, using a preventative model to identify and intervene with students demonstrating poor reading abilities as early as Kindergarten and first grade led to a reduction in student reading difficulties in third grade.

Vellutino et al. (1998) concluded that readers that demonstrated Very Good growth were likely mislabeled as poor readers based upon experiential and instructional deficits whereas those students that showed Very Limited growth were more likely to demonstrate reading difficulties due to cognitive deficits. In addition, the authors found that typical exclusionary criteria, such as those used in typical reading disability identification, failed to adequately identify students who were demonstrating reading difficulties due to experiential and instructional deficits. Similarly, students who demonstrated different response growth rates in intervention were not differentiated by the IQ-achievement discrepancy assessments – the traditional method to identify individuals with a reading disability. These results have been supported across various

studies (Vellutino et al., 1996; Vellutino, Scanlon, & Lyon, 2000; Vellutino, Scanlon, & Zhang, 2007; Vellutino, Scanlon, Zhang, & Schatschneider, 2008).

Vellutino et al. (2007) evaluated longitudinal data in order to assess the effectiveness of using an RTI approach (i.e. early identification and intervention) for identifying students at risk for later academic failure. Vellutino et al. found that Kindergarten students who were identified and received Tier 2 interventions were less likely to demonstrate reading difficulties in first grade. Kindergarten students who did not receive supplemental instruction were more likely to continue to be at risk in first grade. Vellutino et al. also found that measures of growth in early literacy skills were able to distinguish between those students who continued to be at risk for long-term reading difficulties and those students who were able to become independent readers. Thus, the work by Vellutino et al. demonstrates the importance of both early literacy skills and the use of appropriate and effective measurement tools to assess these skills.

Reading Fluency

Reading fluency is the ability to read text aloud with speed and accuracy and it is a critical component of overall reading development (Francis, et al., 2008; National Reading Panel, 2000). According to Perfetti's (1985) theory on verbal efficiency, there is a direct relationship between a reader's rate and his/her comprehension. Thus, students who are fluent are able to focus on comprehending the text whereas less fluent students are more likely to focus on decoding words (Francis, et al., 2008). The theoretical link between fluency and comprehension has received empirical support (see Fuchs, Hosp, & Jenkins, 2001). Given the link between reading fluency and reading comprehension, the

use of measurement tools that assess reading fluency has become increasingly important. One of the earliest attempts at measuring fluency was through CBMs.

CBM and Oral Reading Fluency in RTI

Deno (1985) and colleagues first proposed the idea of using CBM as a tool to monitor the progress of special education students. Reading-CBMs (R-CBM) have received the most empirical attention since its inception (Reschly et al., 2009). The most commonly used form of R-CBM is ORF. ORF measures are used to examine a student's reading fluency by calculating the number of words read correctly per minute (WRCM). ORF scores have been examined thoroughly in the reading literature with research indicating ORF as a predictor of overall reading ability, comprehension, and later reading success (Reschly et al., 2009; Wayman et al., 2007).

Measures of ORF are intended to be *general outcome measures* (GOM; Fuchs & Fuchs, 1999), as opposed to measures that test student mastery of instructional items. GOMs are standardized assessments in which difficulty is held constant across assessments so that long-term goals can be evaluated (Fuchs, Fuchs, & Hamlett, 2007; Stecker, Fuchs, & Fuchs, 2005). There has been approximately 30 years of research supporting the reliability and validity of ORF (Reschly et al., 2009; Shinn, 1998). This research includes empirical support for using ORF data in the classroom to inform and guide instruction. This practice has been demonstrated to lead to increased gains in student performance (Fuchs & Fuchs, 1986).

Some of the key characteristics of ORF measures are that they are designed to be frequently administered, sensitive to changes in the target behavior, relatively

inexpensive, and efficient (Reschly et al., 2009). These characteristics allowed Deno (1985) and his colleagues to provide special education teachers with a new assessment tool that allowed for immediate feedback to be provided to special education teachers regarding the effectiveness of their instruction (Stecker, Fuchs, & Fuchs, 2005). The special education teachers were then able to make informed decisions regarding the need for changes in a student's educational intervention based upon data, with the goal of improving the rate of student learning and increasing overall achievement (Reschly et al., 2009).

The use of ORF to monitor progress was part of a paradigm shift that characterized learning difficulties as problems to be solved rather than inherent intra-child characteristics (Fuchs, Fuchs, McMaster, & Al Otaiba, 2003; Wayman, Wallace, Wiley, Tichá, Espin, 2007). Research has supported the use of the problem-solving approach and, specifically, the effectiveness of using ORF to monitor students' progress (Marston, 1989; Wayman et al., 2007). The use of ORF as a progress monitoring tool has also been demonstrated to have a positive impact on student outcomes (Fuchs & Fuchs, 1986). Based upon early empirical support there was soon a growing body of research that examined the utility, reliability, and validity ORF and other CBMs (Marston, 1989; Olinghouse, Lambert, & Compton, 2006). By the 1990's, CBM research and applications had moved into the general education classroom (Stecker, Fuchs, & Fuchs, 2005).

Along with the increase in the popularity of RTI has come an increase in the scope of the use of ORF (Christ & Silberglitt, 2007; Reschly et al., 2009). ORF measures

are currently used as part of the universal screening and progress monitoring conducted within schools using the RTI model. The ORF data derived from universal screening and progress monitoring are being used as an important part of the decision making process (Christ & Silberglitt, 2007). While data from ORF and other CBMs were originally used in relatively low-stakes decision making (i.e. decisions regarding altering a special education student's academic intervention), they are now used as a key source of data in high-stakes decisions. For example, ORF measures currently provide data that are used to evaluate the effectiveness of instruction, screening, bench-marking, goal setting, and monitoring progress of special education and general education students (Reschly et al., 2009; Stecker, Fuchs, & Fuchs, 2005). Oral reading fluency data have been used to predict later student outcomes on reading measures (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Hosp & Fuchs, 2005) and high-stakes statewide assessments (Buck & Torgesen, 2003; Crawford, Tindal, & Steiber, 2001; Hintze & Silberglitt, 2005; McGlinchey & Hixson, 2004; Silberglitt, Burns, Madyun, & Lail, 2006; Stage & Jacobson, 2001; Vander Meer, Lentz, & Stollar, 2005). In addition, using the dual discrepancy model of learning disability identification (Fuchs & Fuchs, 1998), ORF data can be used as a means to help identify students that demonstrate both an ability level below expectations and a lack of response to research-based interventions. The use of ORF as an assessment tool in such high-stakes decisions has led to increased scrutiny regarding the psychometric characteristics of ORF (Christ & Silberglitt, 2007; Christ & Ardoin, 2009).

Passage Equivalency

One of the assumptions for ORF screening and progress monitoring probes is that each probe is approximately equivalent (Reschly, et al, 2009). This equivalency assumption is critical as it allows for the interpretation of changes in target behavior to be interpreted as a function of changes in a student's skill rather than variability and error associated with the measurement tools (Ardoin, Suldo, Witt, Aldrich, & McDonald, 2005; Jenkins, Zumeta, Dupree, & Johnson, 2005). Thus, it is imperative within the RTI model framework that screening and progress monitoring measures are demonstrably similar. Despite a wealth of research demonstrating the relationship between ORF measures and overall reading abilities (e g. LaBerge & Samuels, 1974; Perfetti, 1992; Shinn, 1989; Fuchs, Fuchs, Hosp & Jenkins, 2001), there is evidence that further statistical equating of passages is required, even when passages demonstrate high correlations and apparent equivalency based upon readability formulas (Francis et al, 2008). Recently there has been a re-examination of the empirical support regarding passage equivalency and thereby the reliability and validity of ORF measures (e.g. Ardoin et al., 2005; Ardoin & Christ, 2009; Betts, Pickart, & Heistad, 2009, Christ & Silberglitt, 2007).

Readability Formulas

The most common means of establishing passage equivalency has been through the use of readability formulas (Ardoin et al., 2005). Readability formulas were developed as a means to identify levels at which students could comprehend text (Balin & Grafstein, 2001). Readability formulas also provide two unique characteristics: an indication of how easy a text is to read; and a quantification of text difficulty (Bailin &

Grafstein). The quantification of text difficulty allows for educators to rank order texts and identify texts of similar difficulty, based upon their readability formula scores.

There have been hundreds of readability formulas developed since the 1920's (Fry, 1989). From these, several readability formulas have become popular, including the Dale-Chall (1948), FOG (Gunning, 1968), Spache (1953), and Fry (1977) formulas. The Dale-Chall formula provides an estimation of reading grade level. In order to obtain the grade level estimate, three 100-word passages from the text of interest are selected. The words within these passages are then compared to a list of 3000 common words, as identified by Dale-Chall (1948). This most updated version of this list was created in 1995 (Micro Power & Light Co., 2000). Bailin and Grafstein (2001) state that the reading grade score is then calculated using the following formula: Reading grade = .16 (% of uncommon words) + .05 (average number of words per sentence). The Dale-Chall formula is designed to identify passages between grades 4 and 8.

Another formula that uses similar techniques is the FOG (Gunning, 1968) readability formula. Accoding to Bailin and Grafstein (2001) the FOG formula is calculated as follows: Reading grade = .4 (average sentence length + percentage of words greater than two syllables). Thus, the FOG formula calculates readability without examining the potential familiarity of the words in the passage.

Spache (1953) sought to develop a readability formula that would be able to more accurately indicate the readability of passages for younger grades, specifically grades 1-3. In addition, Spache used an alternative and reduced word list from the original Dale-Chall list of 3000 words. The new list consisted of 769 words which were, according to

Spache, more indicative of the level of words found in early literacy texts. Thus, grade level was identified by Spache using the following formula: Grade level = .141 (average sentence length per 100 words) + .086 (% of words not on the 769 word list) + .839.

The Fry (1977; 2002) readability formula is computed using a minimum of 3 randomly selected 100-word passages from the text being evaluated. For each 100-word passage the number of syllables and the number of sentences are counted. These numbers are averaged across the three 100-word passages. The average number of syllables and sentences are then plotted on the Fry readability graph (see Fry, 2002). The resulting data point is then used to identify the approximate grade level.

Despite the prevalence and popularity of readability formulas, their validity and reliability has been questioned. Bailin and Grafstein (2001) provide a critique for the tools and theories used to derive readability formulas. The authors describe the scores provided by readability formulas as both "seductive and misleading" (p. 292). The authors state that readability scores are seductive because they lend a sense of scientific and mathematical objectivity and are misleading because this apparent scientific objectivity leads educators to place greater faith in the scores than is warranted by the empirical evidence.

There are hundreds of readability formulas, each with its own criteria used to determine a readability score. However, there are several characteristics that are common across the different formulas. In many readability formulas, the most common variables used in the calculations are semantic difficulty (i.e. vocabulary) and syntactic difficulty (Bailin & Grafstein, 2001; Fry, 2002).

Vocabulary difficulty. Vocabulary difficulty is established via comparison to word lists or using syllable counts (Bailin & Grafstein, 2001). According to Bailin and Grafstein, there are significant limitations to each of these techniques. Word lists are used with the assumption that texts that have a higher number of words that are used more frequently will be easier to read. This assumption is threatened by several important factors. First, the word lists used to evaluate texts were derived at different times in the past. Therefore, some word lists contain words that are no longer frequently used in the modern lexicon. For example, the word "maypole" is on the Dale-Chall 3000 word list but is not likely to be known or used frequently by modern-day school-aged children (Bailin & Grafstein). Similarly, modern-day words such as "download," "internet," and "email" may not be included on word lists created prior to the introduction of these words into the modern lexicon.

Bailin and Grafstein's (2001) second critique of word lists is that the lists are not socio-culturally representative of all individuals. The authors note that some frequently used words by individuals from the inner-city might be distinct from an individual from the suburbs or rural areas. Word lists have been constructed in such a manner so that they are not necessarily representative of the vocabulary from different cultural and regional backgrounds.

The final critique of word lists is that the complexity of the word in use is not considered (Bailin & Grafstein, 2001). For example, individual words may have several different meanings and these meanings may be more or less well known based upon the reader's socio-cultural background and educational experience. In addition, the meaning

of an individual word may only become clear when examining the use of that word within a context, but word lists do not take contextual complexity into account when determining readability.

An alternative to word lists is to examine word complexity which is most frequently calculated by examining the number of syllables within the text. Some readability formulas use the number of syllables per 100 words of texts (e.g. Flesh, Fry), while others use the number of multisyllabic words (e.g. FOG). Across all syllable-based formulas it is assumed that as the number of syllables within the text, and thus the length of the words increases, so does the difficulty of the text (Bailin & Grafstein, 2001).

While examining the number of syllables avoids the socio-cultural vocabulary idiosyncrasies and the variability in word meaning problems inherent to the word list method, there are still significant limitations to this technique (Bailin & Grafstein, 2001). First, monosyllabic words may be more unfamiliar that multisyllabic words. Bailin and Grafstein (p. 289, 2001) note that the words "curr" and "aardvark" are less likely to be familiar to readers than are "reinventing" and "unemployment," yet the latter words are considered more difficult as they contain more syllables than the former. The authors argue that the use of common prefixes and suffixes may actually help increase the comprehension of a text while simultaneous increasing its readability difficulty score using the syllable counting technique.

Syntactic complexity. Syntactic complexity is included in most readability formulas and is typically quantified by examining sentence length (Bailin & Grafstein, 2001). The underlying assumption is that as the length of a sentence increases so does its

difficulty. However, this may not always be accurate and, according to Bailin and Grafstein, longer sentences may actually make them easier to read and be understood than if there were several independent sentences. The authors provide the following sentences as examples (pg. 291, 2001): "I couldn't answer your e-mail." and "There was a power outage." Written separately there is not a clear connection between the two sentences. However, "I couldn't answer your e-mail because there was a power outage." uses the word 'because' to demonstrate the clear relationship between the two ideas and thus make comprehension easier. Readability formulas such as the Dale-Chall (1948), Spache (1953), FOG (Gunning, 1968), and Fry (1977) that use sentence length to calculate readability would score the latter sentence as more difficult despite the potential for it being easier to comprehend than the two separate sentences.

Readability formulas and ORF. Despite these criticisms, the use of readability formulas with ORF has been pervasive. The vast majority of research involving ORF has utilized readability formulas, most typically as a tool to equate passages pulled from students reading curriculum. In addition, readability formulas have been the main source of analysis and data used in the development of several common ORF tools. For example, two commonly utilized and publically available ORF tools are DIBELS (Good & Kaminski, 1996) and AIMSweb (Howe & Shinn, 2002). For both of these tools, the ORF passages that were written were evaluated and revised using readability formulas.

Good and Kaminski (2002) describe the procedures used to identify and equate the DIBELS ORF passages for grades 1 through 3. For each grade the development team wrote passages that were grade appropriate. Each passage was evaluated using several

readability formulas. However there was a high level of variation across readability scores. For example, one first grade passage received readability scores of 2.0, 1.2, 4.1 and 7.1 from four different readability formulas. The authors used only the Spache (1953) readability formula as their final metric for subsequent revisions. Each passage from each grade was evaluated using the Spache formula to determine if the passage fell within the developers' readability formula guidelines. If a passage's score on the Spache formula was too high or low the developers edited the passage in order to reduce or increase its Spache score so that it would reach their criterion.

Good and Kaminski (2002) report that the ORF passages were then rank ordered based upon a composite readability formula score. Based upon this composite score, the probes were divided into three groups: low, middle, and high difficulty. The benchmark and passage monitoring sets were then selected such that each set had one passage from each of these groups. The authors state, "Thus, each benchmark assessment has a first passage representative of the easier third, a second passages representative of the middle third, and a third passage representative of the more difficult third of relative readabilities" (Good & Kaminski, p. 10). Good and Kaminski note that the differences between these passages, based upon the readability composite, were relatively minimal and their goal was to keep the readability of all passages as close as possible.

The progress monitoring probes for AIMSweb also relied heavily upon readability formulas. Howe and Shinn (2002) describe the passage development process for the AIMSweb benchmark and passage monitoring ORF probes. Similar to the DIBELS ORF passage development, the AIMSweb passages were written such that the passages'

readability scores fell within their predetermined readability formula criteria. However, there were several important differences between the techniques that the AIMSweb developers used. Specifically, AIMSweb passages were equated using the Fry (1977) readability formula, the passages were for grades 1-8, and the passage pool was finalized based upon a 'field test' of the passages with students. This field test identified and eliminated passages that demonstrated low alternate form reliability, measured by Pearson's correlation, and the highest levels of variability in mean, standard deviation, and standard error of measurement (SEM). Specifically, passages that showed an alternate forms reliability score of less than .70 and had a mean WRCM of greater than 1.0 SEM outside of the grade-level WRCM mean were eliminated. Thus, in addition to the Fry readability formula, the AIMSweb developers used descriptive statistics to identify those passages that were demonstrating characteristics of non-equivalence when children read the passages.

Based on the apparent methodological and theoretical shortcomings of readability formulas (Bailin & Grafstein, 2001), it is not surprising to see limited support for using readability formulas within the ORF research. This seems appropriate as readability formulas were developed as a means to estimate comprehension difficulty rather than reading fluency (Christ & Ardoin, 2009). Thus, research has generally failed to support readability formulas' ability to predict reading rates because that is not what they were designed to do (Christ & Ardoin, 2009).

An example of the relative weakness of readability formulas ability to identify passage similarity is found in the work of Ardoin, Suldo, Witt, Aldrich, and McDonald

(2005). The authors examined how well several common readability formulas predicted student performance on ORF measures. Six reading passages were analyzed using the following eight readability formulas: Powers-Sumner-Kearl, ([PSK] Powers, Sumner, & Kearl, 1958), FOG (Gunning, 1968), Fry (1977), Spache (1953), Dale-Chall (1948), Flesch-Kincaid ([FK] Flesch, 1948), Forecast (Sticht, 1973, as cited in Ardoin et al.), and SMOG (McLaughlin, 1969), and their components. Of these formulas, five (FOG, Fry, Spache, Dale-Chall, FK) are commonly used in reading research while the remaining formulas (PSK, SMOG, Forecast) are not. When the readability formulas scores were compared to students' WRCM, the authors found that four of the five commonly used readability formulas (Fry, Spache, Dale-Chall, FK) were not significantly related to WRCM. On the other hand, the readability formulas that were not typically used in research (PSK, SMOG, Forecast) all were significantly related to students' WRCM. The authors also found that two components of readability formulas were significantly related to students' WRCM: syllables per 100 words and words not in the Dale-Chall list of 3,000 words. Two other commonly used components (sentence length and Dale-Chall 726 words list) were found not to be significant predictors. Finally, the authors found significant variation and a lack of agreement on passage difficulty rankings between readability formulas; a result that supports previous research on the instability of readability formula estimates (Bruce & Rubin, 1988).

There is research that suggests that readability formulas are able to discriminate between passage levels at a grade-level of analysis (Betts, Pickart, & Heistad, 2009). Betts et al. found that readability formulas they used were able to distinguish passage

difficulties at the between grade-level (i.e. second v. third grade). However, when passages within a grade were examined, the readability formulas were not effective. Specifically, when five intra-grade passages were examined, the authors found that the use of raw scores as a metric for passage equivalence was not supported. That is, the WRCM score derived for ORF passages that is typically used to examine passage equivalency did not accurately identify passages of similar difficulty. In an attempt to identify an alternative method to identify equivalent passages, Betts et al. used horizontal equating using a linear function. Using this statistical equating technique, which adjusts the differences from multiple test forms such that the results are comparable, they were able to statistically equate three of five passages. The authors were also able to identify, using the statistical equating methodology, one passage that was dramatically different from the other passages. Betts et al.'s work highlights the need for research to identify alternative ways of identifying those passages that are truly equivalent, as well as identifying those passages that are non-equivalent and therefore must be altered or removed from an RTI model's progress monitoring system.

ORF Psychometrics

With the increase in its popularity and the importance of the decisions being made, ORF has come under increased scrutiny for its psychometrics. In the recent years there has been a small but important growing body of research that is investigating this topic. Significant shortcomings of ORF passage equivalence, even when there are high levels of correlations across forms and readability indices, have led to the need to use alternative statistical methods to equate passages (Francis et al., 2008). The relatively

few researchers who have investigated the topic have attempted to go beyond readability formulas and highlight some critical potential shortcomings of ORF as well as potential alterative passage equating techniques.

Christ and Silberglitt (2007) highlighted the need to consider the Standard Error of Measurement (SEM) when using ORF in high-stakes decision making. The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) state that scores derived from assessments must include the SEM so that a confidence interval (CI) can be established to estimate the range of likely true scores around an observed score. This idea has been virtually ignored in the ORF literature and in the decision-making process. In order to identify what the SEM for ORF are, Christ and Silberglitt examined eight years of available ORF data for students from grades 1-5. The authors found that the median SEM across grades was 10 WRCM (range 5-15 WRCM). Research has also indicated that SEM and Standard Error of Slope (SEb) can be reduced for both point estimates and trend estimates when passage difficulty is controlled (Hintze & Christ, 2004), and the length of time monitoring progress is increased (Christ, 2006).

Christ (2006) examined ORF data from previously published articles where Standard Error of Estimates (SEE) of ORF growth were provided. With these data Christ was able to estimate the SEb. However, Christ found only 3 articles (from a total of 234 articles) that provided this information. Christ found that 5 weeks of progress monitoring lead to a smaller median SEb (2.21 WRCM) than did 2 weeks of progress monitoring

(median = 9.19 WRCM). The median SEb was reduced to 0.42 when 15 weeks of progress monitoring data were used. Another key finding by Christ was that poorly controlled measurement conditions (e.g. quiet setting, standardized instructions, etc.) can lead to 4 times more error than data collected from optimally controlled conditions.

Generalizability Theory and ORF

The lack of empirical evidence supporting the use of readability formulas as a means to identify equivalent passages has led to a call for research to identify alternative methods to equate passages. To date there have been relatively few attempts to identify means to equate passages using more advanced techniques. Poncy, Skinner, and Axtell (2005) and Christ and Ardoin (2009) provide two of the few articles that propose alternative equating procedures.

One technique that has been proposed to identify equivalent probes is Generalizability Theory (Hintze, Owen, Shapiro, & Daly, 2000). Generalizability Theory, first proposed by Cronbach, Glese, Nanda, and Rajaratnam (1972), is a random sampling theory wherein the dependability of measurement procedures can be analyzed (Marcoulides, 1999). In contrast to Classical Testing Theory's unitary conceptualization of error, Generalizability Theory allows for the examination of several sources of error at once, such as error attributable to measurement items, testing occasions, and test administrators (Marcoulides, 1999; 2000). The analysis of several sources of error allows test developers and researchers to better identify how best to optimize measurement items (Marcoulides, 1999; 2000).

Poncy et al. (2005) used this approach by conducting a generalizability study (G-Study) to identify the percentages of variance that were attributable to student, passages, and error. The authors administered 20 DIBELS ORF passages to 37 third grade students. The authors found that approximately 10% of the variance was attributed to the passages, when all 20 passages were examined. Of the remaining variance, 81% was attributed to the individual and 9% was error. The authors found that the percentage of variance attributed to the passages could be reduced by restricting the number of passages used based upon the average WRCM. When only the passages within a 15 WRCM (n = 14) of the average were used, the percentage of variance attributed to passages decreased to 5.5% and the percentage attributed to the individual increased to 85.5%. When passages within 10 WRCM (n = 11) were used, the variances for passages and individuals were 2% and 89%, respectively. Finally, when passages within 5 WRCM (n = 7) of the average were examined, the percentage of variance attributed to passages and individuals were 1% and 89%, respectively. Thus, Poncy et al. demonstrated that by identifying passages that demonstrate a restricted range of WRCM (i.e. are more similar based upon mean WRCM) the amount of variance attributable to the passages could be reduced from 10% (when all passage are used) to 1% (5 WRCM range). Simultaneously, the amount of variance attributable to the individual increased from 81% to 89%, using the same conditions. Increasing the amount of variance attributable to the individual, as well as reducing the amount of variance attributable to the measurement items, allows for a clearer interpretation of the data (Marcoulides, 1999; 2000). Poncy et al.'s research adds further support to the need to identify those probes that demonstrate equivalency. It

should be noted that Poncy et al. used statistical analyses to come to their conclusions. Other research has used field testing to determine the effects of altering the probe set used.

Field testing was used in two passage equating techniques by Christ and Ardoin (2009). Christ and Ardoin examined four different passage equating techniques (random sampling, Spache [1953] readability formula, mean level of performance, and Euclidean Distance evaluation) to identify those passages that demonstrated equivalency. The passages, derived from third grade reading textbooks, were administered to second and third grade students. For each equating technique, 20 passages were pulled from a larger pool of 50 passages. A random selection of 20 passages was used for the random sampling technique. The Spache readability formula was used to identify those 20 passages that most closely centered around a score of 3.5. Field testing was used to identify the 20 passages used in the mean level of performance (i.e. the 20 passages most closely distributed about the group mean WRCM) and Euclidean Distance evaluation groups. Euclidean Distance refers to the square root of the sum of squared differences for students' WRCM on the passages. Those passages with the smallest (i.e. less variable) mean Euclidean Distance comprised the 20 passages within this group.

Christ and Ardoin (2009) used several G-studies to identify which of the four techniques were best able to identify consistency in oral reading performance. They found that neither of the non-field based techniques (random sampling and Spache [1953] readability formula) provided optimal passage sets. Both techniques poorly controlled for passage as demonstrated by the amount of variance attributed to the passages in the

G-Study. For both second and third grade students, the random selection (10% and 6%, respectively) and Spache readability formula techniques (5% and 4%, respectively) were the poorest performing techniques. The lack of a discrepancy between the random and readability formula techniques contributes to the extant literature that indicates the lack of utility of readability formulas as a tool for identifying equivalent passages. That is, Christ and Ardoin found that there was not a significant difference between randomly selected passages and passages that were purported to be of equal readability according to the Spache formula.

In contrast, the field testing based techniques (mean level of performance and Euclidean Distance) provided more equivalent passage sets. The generalizability studies found that for both the second and third grade students, the percentage of variance from the passages was lowest for the Euclidean Distance technique (1% for both grades) followed by the mean level of performance technique (4% and 2%, respectively). These results lend support to the use of field testing to identify equivalent passages when deriving ORF passages from grade level text.

Two other important findings from the Christ and Ardoin (2009) article are relevant to ORF research. The first important finding is found when the descriptives for all 50 passages are examined. The authors found that for both the second and third grade students, the difference between average performance on the easiest and the hardest passages was 46 WRCM. Thus, passages derived from the same grade level material can lead to significant variability, which would invariably introduce large amounts of error into the decision-making process regarding students' progress and/or benchmark scores

(Christ & Ardoin, 2009). As a comparison, the differences between the WRCM for the middle of the year at-risk and low-risk benchmark scores provided by DIBELS (Good & Kaminski, 1996) are 17 and 26 WRCM for second and third grades, respectively. However, it should be noted that the DIBELS scores are derived from the median of 3 ORF passages whereas Christ and Ardoin are examining each passage individually.

The second important finding was that, despite the significant variation across passages, there were robust levels of alternate form reliability (.92 and .93 for second and third grade, respectively). Christ and Ardoin (2009) state that these high alternate form reliabilities are found because they are limited to rank ordering. That is, the WRCM derived from the passages consistently rank order the students' performance but they do not provide consistent raw WRCM scores. Therefore, the practice of using absolute interpretations of raw scores, as is typically done with benchmarking and progress monitoring, may be significantly flawed (Christ & Ardoin, 2009). These flaws could be critical, particularly when students' raw scores on these passages were as large as 46 WRCM, as was found in this study.

In the Christ and Ardoin (2009) study, the Euclidean Distance equating technique provided the lowest percentage of variance attributable to passages. Those passages, numbering 20, were identified via field testing and subsequent statistical analyses. Ardoin and Christ (2009) used those same 20 passages as part of another study that examined other important passage equating factors. Specifically, the authors evaluated passage equivalency through the examination of the SEE and SEb. Both SEE and SEb are important factors that must be considered within ORF research and school-based

application. The SEE provides an estimation of the range of scores around a student's ORF observed score. Similarly, the SEb provides a range around the slope found across a student's data points. These two factors are important to consider when using ORF data in the decision-making process. SEE's and SEb's that are too large, relative to their respective scores, make interpretation and examination of a students' progress within an intervention difficult, if not impossible (Ardoin & Christ, 2009). A large SEb will also lead to an increased number of false positive and false negative decisions about students' response to an intervention. Increased false positive and false negatives will increase the number of students that are erroneously placed into interventions and, therefore, further tax school districts' limited resources (Ardoin & Christ, 2009).

In addition to the 20 passages identified from the Christ and Ardoin (2009) study, Ardoin and Christ (2009) examined each of these factors for 20 DIBELS and AIMSweb passages. Ardoin and Christ found that the passages equated using the Euclidean Distance method contained the lowest levels of measurement error (SEE = 10.68; SEb = 0.64), followed by the AIMSweb passages (SEE = 11.89; SEb = 0.71) and DIBELS passages (SEE = 15.26; SEb = 0.91). The authors note that the two passage sets that incorporated some level of field testing (Euclidean Distance passages, AIMSWeb), demonstrated lower levels of measurement error than the DIBELS passages, which relied solely upon readability formulas to equate passages. However, while the SEE and SEb for the Euclidean Distance method passages were the lowest of the group examined in this study, the measurement error levels were still large. The authors calculated that the 68% confidence interval (CI) around the slope observed for the Euclidean Distance

method passages was a weekly growth rate of 0.28 – 2.44 WRCM. The large CI found makes accurate interpretation of a student's actual level, growth, and rate of improvement a difficult task.

Optimal Amount of ORF Data

Another area within ORF that requires empirical investigation is the number of probes used during the ORF assessment process. ORF measures are designed such that they are efficient, sensitive to change, and are capable of modeling growth (Fuchs & Fuchs, 1999). The typical procedure used in ORF research is to take the median score of three ORF probes (Shinn, 2002). There is a need for empirical assessment of the difference in variance accounted for when the number of probes used is varied. There have been a few articles that have attempted to address this research limitation.

Poncy et al. (2005) used a decision study (D-study) to examine the SEMs when the number of probes used was varied. D-studies are follow-up analyses of data derived from G-studies within Generalizability theory (Marcoulides, 1999). Analogous to the Spearman-Brown Prophecy Formula used in Classical Test Theory (Marouclides, 2000), a D-study allows for researchers to examine data derived from a G-study and then make recommendations for change in either test items, length, administrators, or whichever facet is of interest to the researcher (Marcoulides, 2000). Thus, a D-study allows for the estimation of the change in the percentages of variance accounted for and error attributable to different facets when those facets are altered. For example, after conducting a G-study using a 50-item test, a researcher may use a D-Study to estimate the percentage of variance accounted for that is gained or lost by adding or reducing the

number of items on the test by examining the resulting generalizability coefficient. The generalizability coefficient is Generalizability Theory's indicator of reliability (Marcoulides, 1999; 2000).

Applying the D-study technique to their G-study data, Poncy et al. found that when one probe was used, the generalizability coefficient was .90 with a SEM of 12 WRCM. When nine probes were administered, the reliability coefficient was .99 with a SEM of 4 WRCM. However, the administration of nine probes would be time consuming and counter to the RTI principle of brief assessments to inform decision making. Thus, the more common and efficient practice of administering three probes was examined. The reliability coefficient and SEM of three probes was .93 and 7 WRCM, respectively. The 7 WRCM SEM for the median of three probes was less than that found by Christ and Silberglitt (2007).

Jenkins, Graff, and Miglioretti (2009) examined how changes in the number of probes administered affected reliably estimates of reading growth using ORF passages. The authors' sample consisted of 41 students with identified learning disabilities and in special education in grades ranging from third through eighth. The authors used ORF passages developed at Vanderbilt University. The authors reported that the passages' difficulty levels were controlled by analyzing them using 20 different readability formulas. The authors did not report which formulas were used or how the results were used to identify which passages met their criteria.

Each student was administered a total of 29 ORF passages over a 10-week period. An overall "true" growth slope was calculated by taking the mean growth rate, across

students, derived from WRCM scores from all 29 ORF passages. Over the 10-week period, the first week served as baseline, which was derived from the average WRCM from four passages. Students then read passages each week for the duration of the study. The number of passages a student read was determined by the study's design. Student growth rates were calculated from ORF passages collected in the following manner: one passage every week; two passages every 2 weeks; three passages every 3 weeks; four passages every 4 weeks; and passages from the first (baseline) and last week only. The authors found that the three passages every 3 weeks (M=1.08) and first and last weeks only (M=1.04) measurement schedules provided growth means that most closely resembled the "true" growth rate estimation (M=1.09).

Jenkins et al. (2009) also examined the effects of using one ORF measure on growth estimation reliability. In the original design, students' ORF baseline was established by taking the average score of four ORF passages. The authors used four passages instead of the customary three passages in an effort to establish a more stable baseline score (Jenkins, et al.). When comparing the number of probes used (1 v. 4), the first probe given at baseline was used in the 1-probe group. The authors found that using one CBM measure instead of four to establish a baseline significantly inflated growth rate estimations. Similarly, when one ORF measure was used at each point of the 10-week design the growth rate estimation was significantly inflated.

Jenkins et al. (2009) used as their outcome measure a "true" reading growth slope which was found by calculating the growth slope derived from 29 ORF probes administered over the 10-week period. Thus, the authors were able to compare the

growth rates calculated from each of the weekly variations (i.e. one every week, two every two weeks, three every three weeks, four every four weeks, and first and last weeks only) to the growth slope using all the measures in order to best identify which was most similar to the "true" rate. The authors' use of the "true" slope as their criterion measure is justified by stating that the use of a pre and post-test design using typical achievement tests would be inappropriate.

Specifically, achievement assessments are inadequate measures of growth over the short-term and are designed to measure relative status rather than individual growth. While the reasons for avoiding using achievement assessments as a criterion measure are valid and logical, the authors' alternative criterion measure of a "true" growth slope is insufficient. One of the critical components of progress monitoring is its ability to provide data that can be used to identify those individuals that are non-responders and are therefore more likely to be at risk for future reading failure (Vellutino et al., 1998; Vellutino, Scanlon, & Zhang, 2007). The 'true' rate criterion proposed by Jenkins et al. may be a limitation as identifying a student's growth rate limits the level of interpretation to that variable. As comprehension is the ultimate goal of reading education, criterion measures that focus on this aspect of reading may be more informative and, as such, should be used as the outcome variable.

Limitations of Previous ORF Research

There are several significant limitations to and unanswered questions within the existing universal screening and progress monitoring research. Throughout the RTI model, data are used to inform the decision making process. Thus, in order for the RTI

model to be successful, the measures used to derive data must be psychometrically sound. In particular, the equivalency of passages is an area that remains elusive. For the most common commercially available universal screening and progress monitoring tools (DIBELS and AIMSweb), the use of readability formulas has been the main source of passage equating. However, there is ample evidence that indicates that readability formulas do little to provide equivalent passages based upon the raw WRCM scores (Francis, et al., 2008). Thus, alternative means to equate passages must be evaluated.

While there have been several articles that indicate positive signs for the use of ORF measures there are significant limitations, beyond readability formulas, that need to be addressed empirically. One of the most striking limitations within the progress monitoring research is the lack of research that uses high-stakes statewide assessment as the outcome variable. Some articles (e.g. Jenkins et al., 2009) use growth rate as outcomes. However, given the importance placed upon statewide assessments by school administrators, there is a need for a better understanding of how progress monitoring data can predict student's outcome on high-stakes statewide assessment. While it is important that the relationship between ORF measures and later ORF reading performance is known, given the high-stakes nature of decisions being made within education, the relationship between ORF measures and statewide assessments needs to be further evaluated.

Several articles have used high-stakes statewide assessments as the outcome variable. However, much of the research that has been conducted using ORF data to predict statewide assessments has several significant limitations of its own. Specifically,

some of the research has not appeared in peer-reviewed journals (Barger, 2003; Buck & Torgesen, 2003; Shaw & Shaw, 2002; Vander Meer, Lentz, Stollar, 2005; Wilson, 2005) or it has a participant pool with a homogenous group or low number of subjects (Barger, 2003); much of the research has used students who have previously been identified and are already receiving intervention or are in special education. The predictive ability of progress monitoring data for student outcomes on high-stakes statewide testing, above and beyond information provided by universal screening data, has yet to be answered.

Another area in the ORF progress monitoring research that requires further evaluation involves the number of progress monitoring data probes and overall number of data points. Specifically, the number of probes and data points that are needed to provide the most amount of predictive ability while remaining efficient requires further clarification. For example, Jenkins et al. (2009) indicated that the number of progress monitoring assessments could be reduced. While their research indicates a reduction in the number of assessment time points will not negatively impact the estimation of the true slope, it ignores the basic tenet that progress monitoring is done to inform the decision making process and quickly identify non-responders. The authors' assertion contradicts other research that indicates that more data points leads to greater ability to accurately identify student abilities (Christ, 2006).

Another limitation within the ORF progress monitoring literature is the combination of student data across grades. For example, Jenkins et al. (2009) found a mean growth rate of 1.09 across all 29 passage administrations. This mean growth rate was used as the standard against which their different time administration rates were

measured. However, this growth rate is not representative of the growth rate across all the groups (Fuchs, Fuchs, Hamlett, & Germann, 1993). Similarly, Christ and Ardoin (2009) combined students' data from second and third grade students into a single analysis. The combining of second and third grade students' WRCM scores ignores existing research that indicates students' reading ability and their ORF growth rates are not uniform across grades (Deno, Fuchs, Marston, & Shinn, 2001). By combining these students' scores, differences between the predictive accuracy of later reading success may be obscured.

Another limitation within the ORF literature involves the lack of standardization of curriculum passages used. Some research has used readability formulas to identify ORF passages pulled from student grade-level texts (Francis et al., 2008). However, as there are significant limitations of readability formulas (Ardoin et al., 2005; Betts et al., 2009; Christ & Ardoin, 2009), the stability of these passages are in question. Some research, such as Jenkins et al. (2009), identified the passages used in their research through teacher recommendation of passages at reading level, while other research (Christ & Ardoin, 2009) used passages derived from third grade text to evaluate both second and third grade students. The lack of standardization and the lack of a match between student's grade level and the ORF passage level may introduce unaccounted error into the research designs.

A final limitation to the current ORF research is the relatively low number of participants within many of the studies. The number of participants is reduced even further when the data are not collapsed across grades, as mentioned in the first limitation.

The relatively low number of students used in the existing ORF research limits the ability to generalize the findings beyond the populations used.

Research Questions

The focus of these research questions was on limitations within the existing literature. Specifically, this study attempts to add to the ORF literature in two areas: ORF passage equivalency and ORF's predictive accuracy of a high-stakes statewide assessment. In order to address these needs, two of the most popular CBM assessment tools, DIBELS Next (Good and Kaminski, 2010) and AIMSweb, were utilized. These assessments were chosen because of their frequent use in both research and school district settings. In addition, passages derived from a school district's curriculum (SDC) were used in the analyses.

The focus of research questions 1a – 1b was on the accuracy of readability formulas and the relationship between readability formula rankings and student WRCM. The focus of research question two was on the distribution of ORF passage variability between students, passages, and error. The focus of research questions 3a – 3b was on the limitations related to progress monitoring. Finally, the focus of research question four was to examine the optimal number of probes to use. The research questions are as follows:

- Research Question 1a: To what extent do the Flesch, FOG, Powers, SMOG, Forcast, Fry, Spache (3rd grade only), and Dale-Chall (5th grade) readability formulas provide similar readability scores for ORF passage from the DIBELS Next, AIMSweb, and SDC passage sets?

- Research Question 1b: To what extent are readability formula scores related to student's WRCM for ORF passages from the DIBELS Next, AIMSweb, and SDC passage sets?
- Research Question 2: What percentage of variance is attributable to item (i.e. passage), the individual, and error for third and fifth grade ORF passage sets from DIBELS Next, AIMSweb and SDC? G-Theory was used in order to evaluate passage equivalency.
- Research Question 3a: To what extent does the addition of the slope from six weeks of ORF progress monitoring data explain additional variance in predicting later reading success, as measured by the California Standards Test-English Language Arts (CST-ELA) beyond universal screening ORF data? In addition, does any one of the ORF passage sets (i.e. DIBELS Next, AIMSweb, SDC) predict CST-ELA scores better or worse than the other passage sets?
- Research Question 3b: Do bi-weekly or first three weeks only progress monitoring models provide a more parsimonious alternative to the 6-week model when explaining variance in predicting CST-ELA outcomes?
- Research Question 4: To what degree is the amount of variance accounted for different between a one-randomly-chosen-probe approach score and a median-of-three-probe approach score? For this question, the data for the one-randomly-chosen approach are derived by randomly choosing one of the three weekly ORF probes as the score for that week. In addition, does any one of the ORF passage sets (i.e. DIBELS Next, AIMSweb, and SDC) predict CST-ELA scores better or

worse than the other passage sets, when using the one-randomly-chosen-probe approach score?

Method

Procedure

Parental consent was obtained for 114 students from two schools within an urban school district in Southern California. Prior to participation in the study, student assent was collected. Two of the students did not provide assent giving a final total of 112 students: 49 third-grade (46.9% male) and 63 fifth-grade students (44.4% male). Student demographic data were collected by school staff and provided to the author by school administrators. Students attended a district where approximately 81.3% of students were receiving free or reduced lunch. Oral reading fluency data were collected individually by trained test administrators. As per Christ's (2006) recommendations, efforts were made to collect all data in a quiet environment and to follow standardization procedures. Data were collected either in a quiet corner of the student's classroom or outside the student's classroom.

In order to answer research questions 1 and 2, students were randomly assigned to read from one of the three passage sets (DIBELS Next, AIMSweb, SDC). For the DIBELS and AIMSweb passages the respective standardized directions were administered to each student. For the SDC passage set the AIMSweb directions were given to each student. In order to reduce maturation effects as much as possible (Fuchs & Fuchs, 1993) while simultaneously avoiding fatigue effects, efforts were made for students to read all 20 passages over a 2-day period (i.e. 10 passages per day). The

majority of student data (81%) were collected within two days and a total of 89% were collected within four days. The remaining 11% of student data were collected between 6-11 days due to factors such as scheduling conflicts, school functions, and absences.

In order to answer research questions 3 and 4, progress monitoring ORF passages were administered to students. The same student participants used to answer research questions 1 and 2 were used to answer research questions 3 and 4. Each student was randomly assigned to read one of the passage sets that he or she did not read previously. For example, if a student read the DIBELS Next ORF passages as part of research questions 1 - 2, then that student was randomly assigned to read either the AIMSweb or SDC passages. By ensuring that no student read the same ORF passage set twice the potential for practice effects, a source of unwanted error, was reduced.

For each student there were weekly assessments during a 6-week progress monitoring period. During these weekly assessments each student read three ORF passages from his or her randomly assigned passage set, giving a total of 18 passages read over the 6-week period. For the DIBELS Next and AIMSweb passages the respective standardized administration instructions were administered to each student. Similar to the procedure for research questions one and two, students reading the SDC passage set were given the AIMSWeb directions.

Administrator Training and Interobserver Reliability

Each test administrator was trained in the administration and standardization for each of the passage sets. Three of the test administrators, including the author, had extensive experience in administering ORF passages in previous research. A fourth test

administrator was trained by the author and was observed on several occasions to ensure procedural integrity. Before data collection began, scores collected by the test administrators were compared to scores collected by the first author in order to ensure accuracy.

In order to determine interobserver reliability the percentage of agreements were divided by the number of agreements plus disagreements (Hintz, 2005). Based on a total of 34 IRR observations the interobserver reliability was 98.33%.

Materials

The ORF passages used in this study were derived from three sources: DIBELS Next, AIMSweb, and the school district's grade-level curriculum.

DIBELS Next. DIBELS Next (Good & Kaminski, 2010) is a set of literacy measures. The DIBELS Next ORF measure is an individually administered standardized assessment of reading accuracy and fluency using grade-level text. The DIBELS Next ORF assessment uses the same principles and procedures of other CBM (see Shinn, 1989). DIBELS Next ORF probes are 1-minute long fluency measures. Students read aloud three grade-level reading passages. The WRCM score is calculated by subtracting errors and omissions from the total number of words read. Misread words or hesitations of greater than three seconds are considered errors, while self-corrections, within three seconds, are considered accurate. The median score of the three passages is the student's oral reading fluency score. The DIBELS Next passage set consists of 21 ORF passages, for each appropriate grade-level, provided by the DIBELS Next developers. For the

progress monitoring portion of this study only the first 18 DIBELS Next passages were used in the analysis and for the passage equivalency section all 21 passages were used.

AIMSweb. The AIMSweb R-CBM measure is an individually administered standardized oral reading fluency measure (Howe & Shinn, 2002). AIMSweb measures use the same principles and procedures of other CBM (see Shinn, 1989). AIMSweb R-CBM probes are 1-minute long fluency measures. Students read aloud three grade-level reading passages. The WRCM score is calculated by subtracting errors from total words read. Misread words or hesitations of greater than three seconds are considered errors, while self-corrections, within three seconds, are considered accurate. The median score of the three passages is the student's oral reading fluency score. Fluency scores are collected at three times during the school year (fall, winter, and spring). The AIMSweb passage set consists of the first 20 passages for each appropriate grade-level retrieved from the AIMSweb website. For the progress monitoring portion of this study only the first 18 AIMSweb passages are used in the analysis and for the passage equivalency section all 20 passages were used.

School District's Curriculum Passages. A passage set derived from a school district's curriculum were used as part of these analyses. English Language Arts books from the *Open Court* curriculum were used as sources for the SDC passages in both 3rd and 5th grade. In order to limit the potential for students finishing a passage before the one minute had elapsed, only passages with a minimum of 200 words were used for this study. This passage set was included along with the other commercially available sets (DIBELS Next and AIMSweb) due to the prevalence of similar passage sets being used

in research and within the school setting. There have been several articles that have included some form of a non-standardized passage set (e.g. Ardoin & Christ, 2009; Hintze & Christ, 2004). Many curricula used in schools now provide their own ORF passages. For the purposes of this research, text-based curriculum passages were retrieved from the grade-level curriculum being used within the school district. Similar to the other ORF measures, the WRCM score is calculated by subtracting errors and omissions from the total number of words read. Misread words or hesitations of greater than three seconds are considered errors, while self-corrections, within three seconds, are considered accurate. The median score of the three passages is the student's oral reading fluency score. The SDC passage set consists of the first 20 passages for each appropriate grade-level retrieved from the *Open Court* curriculum grade-level books. For the progress monitoring portion of this study only the first 18 school curriculum passages are used in the analysis and for the passage equivalency section all 20 passages were used.

Readability Formulas. All three of the passage sets were analyzed using the *Readability Calculations* (Micro Power & Light Co., 2000) program. This program provides multiple readability estimates for a passage. Each passage used in this study was entered into the program and the following readability formulas were calculated: Flesch, FOG (Gunning, 1968), Powers, SMOG, Forcast, Fry (1977), Spache (3^{rd} grade only; 1953), and Dale-Chall (5^{th} grade only; 1948). The Spache formula was used with the 3^{rd} grade passages only because the formula is designed to be used in lower elementary grade level text only (Spache, 1953; Micro Power & Light Co., 2000). Similarly, the Dale-Chall is used with the 5^{th} grade passages only because it is designed

to be used with passages at a 4th grade or higher level of difficulty (Micro Power & Light Co., 2000). Unlike the other formulas that produce numeric grade estimations (e.g. 3.2), the Dale-Chall formula gives broader grade-level estimations (early 5th). In order to include the Dale-Chall output in these analyses the estimation was quantified. Using a 5th grade passage as an example, the Dale-Chall label and quantification were as follows: Early 5th = 5.1; Early-to-mid 5th = 5.3; Mid 5th = 5.5; Mid-to-upper 5th = 5.7; Upper 5th = 5.9. This quantification process was developed by the author in order to include the Dale-Chall in the analyses.

CST-ELA. The CST-ELA served as the main dependant variable for the predictive component of this study. The CSTs are the high-stakes statewide assessment administered to all school children in California, beginning with the second grade (CDE, 2009). The CSTs are a criterion referenced assessment that are used to measure student's mastery of grade-level content in several areas. The English Language Arts (ELA) section of the CSTs assesses students' English language abilities based upon grade-level curriculum and minimum proficiency standards. The test content validity was established via item review by experts in English Language Arts (CDE, 2009). Convergent validity was established by comparing ELA to California Achievement Test - Sixth Edition (CAT/6) Reading and Language tests. Correlations between the second and third grade CST-ELA and the CAT/6 Reading were .77 and .77 and CAT/6 Language .76 and .75, respectively (CDE, 2009). Reliability scores for the CST-ELA were also strong: $\alpha = .94$ (2nd grade); .93 (3rd grade; CDE, 2009).

Each student's CST raw score is converted to a scale score that falls into one of 5 categories: Far Below Basic, Below Basic, Basic, Proficient, and Advanced. The California Department of Education (2009) states that the educational goal for every student is to reach at least the Proficient level on the CST assessments. Scores on the CST-ELA and Mathematics assessments are used as part of the formula to measure school and district Academic Performance Index. In addition, the CST-ELA and Mathematics tests are used to calculate Adequate Yearly Progress scores, which are used to evaluate progress towards NCLB goals of student academic proficiency.

Results

Descriptive data for CST-ELA scores and Winter ORF screening scores are presented in Table 1. Descriptive data for readability formula scores are presented in Tables 2 – 3. Data for 3rd and 5th grade students are presented separately.

Research Question One

For research question one, each of the 21 DIBELS Next, and 20 AIMSweb, and SDC ORF passages were analyzed using the Flesch, FOG, Powers, SMOG, Forcast, Fry, Spache (3rd grade only), and Dale-Chall (5th grade only) readability scores. The readability scores were calculated using the *Readability Calculations* program (Micro Power & Light Co., 2000). For those formulas that compare passage words to a list (e.g. Dale-Chall) to calculate passage difficulty the most up-to-date list available was used.

In order to answer research question 1a, a nonparametric statistic must be used as there cannot be an assumption of equal intervals between grade levels of passages (Ardoin et al., 2005). The Kendall's coefficient of concordance (Kendall's) is a

nonparametric measure of correlation used with ordinal data in order to obtain a correlation between multiple (i.e. greater than 2) sets of ranks (Sheskin, 2007; Siegel & Castellan, 1988). The measure of agreement between sets provided by the Kendall's statistic allows for the examination of agreement between rankings. The Kendall's scores (represented by *W)* can range from 0 to 1, where a score of 1 indicates complete agreement across rankings and a score of 0 indicates no pattern of agreement (Sheskin, 2007). The test of significance for *W* can be completed using chi-square distribution tables when the following formula is used:

$$\chi^2 = m\,(n - 1)\,W$$

Where:

m = number of sets

n = number of objects (i.e. probes)

In order to answer the first research question, the relationship between the seven readability formulas (Flesch, FOG, Powers, SMOG, Forcast, Fry, Spache (3^{rd} grade only), Dale-Chall (5^{th} grade only)) was calculated using the Kendall's statistic. The Kendall's test was conducted for each passage type in both 3^{rd} and 5^{th} grade. As shown in Table 4, the resulting χ^2 statistics indicate that for each grade (3^{rd} and 5^{th}) and for each passage type (DIBELS Next, AIMSWeb, SDC) there was a significant association among the ranking of the seven readability formulas.

To further examine these relationships, the level of association between individual readability formulas and a student's WRCM were evaluated. For each student, a ranking of passage difficulty was created based upon the WRCM measured. The passage with

the lowest WRCM was ranked the most difficult. The passage with the next lowest WRCM was ranked second most difficult and so on until the passage with the highest WRCM was ranked the easiest. Ties in WRCM rankings were handled the same as were the ties in readability formula rankings (see Sheskin, 2007). This process was repeated for each student.

In order to answer research question 1b, Kendall's tau (τ), a non-parametric measure of association, was then calculated between each readability formula's ranking and the passage difficulty rankings as measured by the WRCM for each student. The resulting Kendall's taus were then used to in a Friedman Two-way Analysis of Variance by Ranks. The Friedman, a non-parametric alternative to a classical Analysis of Variance (Sheskin, 2007), was then employed to assess for main effects. For the DIBELS Next passage set there was a significant main effect for the 3rd grade, $\chi^2(6, N = 19) = 25.02$, $p < .001$, and 5th grade, $\chi^2(6, N = 25) = 60.90$, $p < .001$, passages. For the AIMSWeb passage set there was a significant main effect for the 5th grade passages, $\chi^2(6, N = 25) = 46.31$, $p < .001$ but not for the 3rd grade passages, $\chi^2(6, N = 18) = 3.20$, $p = .783$. For the SDC passage set there was a significant main effect for the 3rd grade, $\chi^2(6, N = 11) = 34.55$, $p < .001$, and 5th grade, $\chi^2(6, N = 13) = 47.23$, $p < .001$, passages. Statistical significance on the Friedman indicates that there is at least one readability formula that has a relationship with WRCM that is significantly different from at least one other formula.

Sheskin (2007) recommends using Wilcoxon matched-pairs signed-ranks test (Wilcoxon) as a nonparametric follow-up pairwise comparison when a significant main

effect is found using the Friedman test. For this research, the Wilcoxon test allows for the examination of whether a readability formula is significantly better or worse than another readability formula in predicting WRCM rankings. (Ardoin, et al., 2005). For each significant Friedman main effect, the Wilcoxon follow-up was conducted with a p-value adjusted to .001 to account for the multiple comparisons. The results of follow-up assessments are displayed in Tables 5 – 9. For the DIBELS Next 3rd grade passage set the Fry readability formula was significantly different from the FOG and SMOG formulas. All other differences were non-significant. Although there was a main effect found for the SDC passage set using a p-value of .05, there were no significant differences between readability formula ranks when using the adjusted p-value of .001 was used.

For the 5th grade version of the passages there were a higher number of significant differences. For the DIBELS Next passages the Flesch formula was significantly different from all formulas except for the Powers formula. The Powers formula was significantly different from all other formulas except for the Flesch formula. For the AIMSWeb passages the Flesch formula was significantly different from the FOG, SMOG, & Fry formulas. The Dale-Chall was significantly different from all formulas with the exception of the Flesch. For the SDC passages the Dale-Chall was significantly different from the Flesh and FOG formulas while all other comparisons were non-significant.

Research Question Two

In order to answer research question two, a Repeated Measures Analysis of Variance (ANOVA) was conducted. The ANOVA output was used to complete analysis for a one-facet G-study. In the one-facet design, the model is written as follows (Marcoulides, 2000):

$$X_{pi} = \mu + \mu_p + \mu_i + \mu_{pi,e}$$

Where:

μ = Grand Mean

μ_p = person effect

μ_p = item effect

$\mu_{pi,e}$ = residual effect (i. e. error)

The ANOVA was calculated using the data derived from the ORF measures. Using a Generalizability Theory framework, variance components for the person (i.e. student), item (i.e. passages), and residual (i.e. error) were calculated using the appropriate formulas (Marcoulides, 2000). In addition, a generalizability coefficient (G-coefficient) using relative error was calculated. G-coefficients are used as an indication of the dependability of a measurement procedure and range from 0 to 1.0 with higher scores indicating higher reliability (Marcoulides, 2000). The specific formulas used in the calculations are as follows:

Person variance scores:

$$\sigma^2_p = (MS_p - MS_{pi,e}) \; / \; n_i$$

Item variance scores:

$$\sigma^2_i = (MS_i - MS_{pi,e}) \; / \; n_p$$

The error variance component ($\sigma^2_{pi,e}$) is equal to the Mean Square of the residual produced by the repeated measures ANOVA. G-coefficients using relative error were then calculated using the following formula:

$$E\rho^2_\delta = \sigma^2_p / (\sigma^2_p + \sigma^2_\delta)$$

Where: $\sigma^2_\delta = \sigma^2_{pi,e} / n_i$.

For 3rd grade students who read the DIBELS Next passages the amount of variance attributable to the students was 89.66%, to the passages was 1.74%, and to error was 8.59%. The G-coefficient was .995 for 3rd grade students reading DIBELS Next passages. For 3rd grade students who read the AIMSWeb passages the amount of variance attributable to the students was 95.52%, to the passages was 0.73%, and to error was 3.74%. The G-coefficient was .998 for 3rd grade students reading AIMSWeb passages. For 3rd grade students reading the SDC passages the amount of variance attributable to the students was 88.53%, to the passages was 4.34%, and to error was 7.12%. The G-coefficient was .996 for 3rd grade students reading the SDC passages.

For 5th grade students who read the DIBELS Next passages the amount of variance attributable to the students was 92.13%, to the passages was 2.67%, and to error was 5.19%. The G-coefficient was .997 for 5th grade students reading DIBELS Next passages. For 5th grade students who read the AIMSWeb passages the amount of variance attributable to the students was 88.8%, to the passages was 3.2%, and to error was 8%. The G-coefficient was .995 for 5th grade students reading AIMSWeb passages. For 5th grade students reading the SDC passages, the amount of variance attributable to the students was 82.47%, to the passages was 10.75%, and to error was 6.77%. The G-

coefficient was .996 for 5th grade students reading the school district's curriculum passages. Taken together, the reliability and variance attributable to students are high for both the 3rd and 5th grade students. These results exceed those found by Poncy, Skinner and Axtell (2005) and provide strong evidence for reliability of all three of the passage sets.

Research Question Three

In order to answer research questions 3a – 3b, a multiple regression design was used. Unlike previous research which regressed variables upon reading growth rates (e.g. Jenkins et al., 2009) the outcome variable in this study was the standard score on a high-stakes statewide assessment (i.e. CST-ELA). The predictor variables for the regression equation varied in order to answer the specifics of each question. For both 3a and 3b, each set of regression equations was calculated separately using DIBELS Next, AIMSweb, and the school district's curriculum passage set.

For question 3a, the first model was a simple linear regression using students' winter ORF benchmark scores to predict CST-ELA outcomes. Model 2 added students' ORF slope, derived from six weeks of progress monitoring data, to Model 1. The slope was derived by calculating a linear regression through each of the six data points for each child. This design was used in order to examine what, if any, additional variance in CST scores can be accounted for by six weeks of ORF progress monitoring data above winter ORF universal screening scores.

As seen in Tables 10 - 12, for 3rd grade students, for the students who read the DIBELS Next passages the winter ORF screening was a significant predictor of CST-

ELA scores. However, the winter ORF screening score for the students who read the AIMSWeb and SDC passages was not a significant predictor of CST-ELA outcomes. Thus in Model 1, before student progress monitoring data were included, there were group level differences between the predictive relationship of winter ORF screening and CST-ELA outcomes. This group-level difference is unexpected as all students were randomly assigned to one of the three reading passage set groups. The addition of progress monitoring slope did not significantly improve the model for any of the third grade passage groups.

The same regression model was repeated using data from students in the 5th grade. For these students, there was a significant association between winter ORF screening and CST-ELA scores for all three groups. The addition of progress monitoring slope did not significantly improve the model for any of the 5th grade passage groups. The lack of additional variance accounted for by adding slope replicates other research. For example, Schatschneider, Wagner, and Crawford (2008) found that the addition of ORF slope did not provide additional information beyond a single time point assessment, for first grade students.

For research question 3b, the progress monitoring data collected across the 6-week period were analyzed to determine the most effective and efficient use of resources. In order to examine this, several regression equations were analyzed using the ORF score from each of the progress monitoring weeks. The first regression equation, which includes all 6 data points, was labeled the complete model. Subsequent regression equations, labeled reduced models, removed specific weeks. The bi-weekly model

included data from weeks 2, 4, and 6 only. The first three weeks model included data from weeks 1, 2, and 3 only. Each reduced model was then compared to the complete model to compare the change in variance accounted for by both the complete and reduced models (Agresti & Finlay, 1997).

The regression equation outcomes for the complete and reduced models are presented in Tables 13 - 18. For 3rd grade students, the complete models for the DIBELS, AIMSWeb and SDC passage type were not significant. This result indicates that the six weeks of progress monitoring data did not significantly predict CST-ELA outcomes for any of the passage sets. For the DIBELS Next passage set, the first-three-weeks-only model was significant, $F(3, 8) = 5.30$, $p = .026$, although the ΔR^2 statistic between the complete and reduced models was not significant, $\Delta R^2 = -.16$, $p = .319$. This result, likely due to a change in degrees of freedom, indicates that the reduced first-three-weeks-only model may be a more parsimonious model than the complete.

For 5th grade students, the complete model using DIBELS Next passages was significant while the complete models using AIMSWeb and SDC passages were not. Further examination of the DIBELS Next results found that there was not a significant difference for the bi-weekly, $\Delta R^2 = -.177$, $p = .205$, or the first three weeks only $\Delta R^2 = -.104$, $p = .401$ reduced models. Both of these models were significant and therefore are parsimonious alternatives to the complete model. Further examination of the AIMSWeb results found that both of the reduced models were significant. Although neither the ΔR^2 for the bi-weekly model, $\Delta R^2 = -.097$, $p = .511$, nor the first three weeks model, $\Delta R^2 = -.083$, $p = .527$, were significant, their respective overall models were significant,

indicating that the reduced models provide a more parsimonious alternative than the complete model. Further examination of the SDC passages found that both of the reduced models were significant. Although neither the ΔR^2 statistics for the bi-weekly model, ΔR^2 = -.07, p = .617, nor the first three weeks model, ΔR^2 = -.181, p = .251, were significant, their respective overall models were significant. Thus, for 5th grade students, both of the reduced models were significant for each of the three passages types indicating that the reduced models are more parsimonious alternatives. The significance of the models, despite non-significant ΔR^2, is likely due to a reduction in degrees of freedom when calculating the F-statistic.

Research Question Four

For research question four, the analyses from research question three were re-conducted. However, instead of using the median score from three ORF passages, the analyses were conducted using one randomly selected progress monitoring probe as the source of weekly data. The same complete and reduced model comparison used to answer research question three were replicated to answer research question four.

Thus, for the replication of research question 3a, there was no change in the first model which was a simple linear regression using students' winter ORF benchmark scores to predict CST-ELA outcomes. However, the slopes added to the second models were calculated using one randomly selected passage score from each week instead of the weekly median scores. The slope was calculated the same was as previously described with the following exception: the data used to calculate slope were derived from one randomly-selected data point per week instead of the median of three. The results for

questions four are presented in Tables 19-21. For 3rd grade students, there was a significant association between winter ORF screening and CST-ELA scores for 3rd grade DIBELS Next passages but not for AIMSWeb and SDC passages. The addition of progress monitoring slope did not significantly improve the model for any of the third grade passage groups. For the DIBELS Next group, despite lacking a significant change in variance explained, $\Delta R^2 = .001$, $p = .918$, the addition of the progress monitoring slope to the regression equation, and subsequent loss of a degree of freedom, made the model non-significant.

The same regression model was repeated using data collected from students in the 5th grade. For these students, there was a significant association between winter ORF screening and CST-ELA scores for all three groups. The addition of progress monitoring slope did not significantly improve the model for any of the 5th grade passage groups.

The final portion of research question four involved the replication of question 3b in which the same previously described procedures for using one-randomly-chosen-probe approach were utilized. The complete and reduced models are in Tables 22 - 27. For 3rd grade students, the complete models for the DIBELS and SDC passage sets were not significant while the complete model for the AIMSWeb passage group was significant. For both the DIBELS and SDC passage sets, neither the bi-weekly nor the first three weeks models were significant improvements ($p > .05$) over the complete models. Further examination of the AIMSWeb passage group found differing results for the bi-weekly and first-three-weeks-only reduced models. The ΔR^2 for the bi-weekly reduced model was non-significant, $\Delta R^2 = -.101$, $p = .178$, and the overall model was significant

indicating that the reduced model is a more parsimonious alternative to the complete model. The ΔR^2 for the first three weeks reduced model was significant, ΔR^2 = -.592, p = .011, and the subsequent model was non-significant. Thus, the first-three-weeks-only reduced model was not a more parsimonious alternative to the complete model.

For 5th grade students the complete model for the DIBELS Next, AIMSWeb, and SDC passage groups were significant. Further examination of the DIBELS Next results found that there was not a significant difference for the bi-weekly (ΔR^2 = -.175, p = .218) or the first-three-weeks-only (ΔR^2 = -.20, p = .176) reduced models. Each of the reduced models were significant indicating that either one is a more parsimonious alternative to the complete model for the DIBELS Next group. Further examination of the AIMSWeb results found that there was not a significant difference for the bi-weekly (ΔR^2 = -.29, p = .050), or the first-three-weeks-only (ΔR^2= -.139, p = .208) reduced models. Each of the reduced models were significant indicating that either one is a more parsimonious alternative to the complete model for the AIMSWeb. Further examination of the SDC results found that there was not a significant difference for the bi-weekly (ΔR^2 = -.219, p = .128), or the first-three-weeks-only (ΔR^2 = -.027, p = .825) reduced models. Each of the reduced models were significant indicating that either one is a more parsimonious alternative to the complete model for school district curriculum group.

Discussion

One of the core components in an RTI model is the assessment of students' current level of skill, either as part of whole-class screening or as part of a more frequent progress monitoring schedule. The quality of the decisions educators can make is

directly related to the quality of data derived from these assessments. In the elementary school years one of the most common and empirically supported forms of CBM is ORF. The general purpose of this study was to evaluate the quality of three types of ORF assessments. More specifically, there were two areas of focus in this research. First, the reliability of readability formulas used to calculate ORF passages was examined by research questions one and two. Second, the utility of using ORF progress monitoring passages to predict a high-stakes state-wide assessment was examined by research questions three and four. For all research questions data were collected from students in two grades (3rd and 5th grade) using three different types of ORF passages (DIBELS Next, AIMSWeb, SDC).

Passage Equivalency and Readability Formulas

Examination of the equivalency of readability formulas, using the Friedman test, a nonparametric version of ANOVA, provided mixed results. There were significant differences in rankings for the DIBELS Next and SDC passages but not the AIMSWeb for 3rd grade students. There were significant differences for all three passage types for 5th grade students. Follow-up pairwise comparisons found that differences in the third grade were minimal. For the DIBELS Next passages only two of the associations were significant different (Spache v. SMOG and Spache vs. FOG). For the SDC there were no significant differences once the p-value had been adjusted for multiple comparisons. Thus, for 3rd grade passages as a whole, the readability formulas ranked the difficulty of the passages similarly (with the exception of the two differences on the DIBELS Next). However, as presented in Tables 2 and 3, the mean grade-level scores provided by the

different formulas show a wide range. Therefore, with the exception of the two Spache differences for the DIBELS Next passages, the other readability formulas demonstrate stability in the ranking of the individual passages within the parameters of their own formula. For example, the mean passage grade-level ranking for the 3rd grade DIBELS Next passages using the Spache formula was 3.79 whereas the mean for the FORCAST was 8.76. However, there was not a significant difference between the rankings of these two readability formulas. Thus, for the third grade passages, although the grade-level score of an individual passage may be dramatically different across readability formulas, the difficulty of a set of passages can be ranked relatively equally regardless of the type of formula used.

The results for the 5th grade passage sets are less clear. The Friedman test was significant for the DIBELS Next, AIMSWeb, and SDC passages. For the DIBELS Next set, two formulas, Flesch and Powers, had significantly different grade-level rankings from all the other formulas (though there was not a significant difference between the two). Thus, the Flesch and Powers ranked passage difficulty in a similar manner, yet significantly different than all the other formulas.

For the AIMSWeb passages the Flesch formula was significantly different from three other formulas (FOG, SMOG, Fry) while the Dale-Chall was significantly different from all the other formulas. This discrepancy was seen, albeit on a smaller scale, in the SDC passages. For those passages the only observed significant differences were between the Dale-Chall and two other formulas (Flesch and FOG). The Dale-Chall formula results are particularly intriguing because it is a formula that is designed to be

used with passages from upper-elementary and higher. However, these results may have been influenced by the quantification process required to analyze the Dale-Chall results. It is possible that the quantification may have introduced error into the subsequent analyses.

Similar to the 3rd grade passages, there was a large range of mean passage difficulty ranks between formulas for the 5th grade passages. A few formulas, particularly the Flesch and Dale-Chall, appeared to have different rankings from many of the other formulas. However, there was also a fair amount of agreement in passage rankings by set making interpretation or recommendation on which readability formula is best suited for 5th grade passages difficult.

Passage Variability Using Generalizability Theory

G-Theory was used as a way to examine the reliability of ORF passages. Replicating and expanding upon results from previous research (e.g. Poncy, Skinner, Axtell, 2005), the result of this study provided strong evidence that the vast majority of the observed variance was attributable to students and not to the passages or to error. This effect was observed across 3rd and 5th grade and all passage types. The G-coefficient, a measure of reliability, was greater than .99 for all passage types at all grades, indicating a high amount of reliability in passage scores. There were high levels of observed variability attributable to the students across all settings (ranging from 82.47 – 95.52%) coupled with the low passage variability (ranging from 0.73 – 10.75%) with the commercial passage sets (DIBELS Next and AIMSWeb) outperforming the SDC passage set. These results provide strong evidence to support the idea that student scores

are attributable to differences in student abilities and not differences in the passages or random error. In addition, based upon the slightly higher levels of error attributable and lower levels attributable to passages, the commercial passage sets (DIBELS Next and AIMSWeb) appear to be a more reliable alternative to the SDC passages.

Taken together, the results from research questions one and two allow for some interpretation of ORF passages. In general, educators can select standardized passages and anticipate ORF scores that are highly reliable and are reflective of student ability, rather than error attributable to passages. These strong psychometric properties are found across passages despite the wide range of readability formula scores. The utility of readability formula scores as a means to accurately equate passages needs to be reconsidered. At a minimum, readability formulas should be used as a first step in a passage equating procedure. The wide range of scores found in this study indicates that passage readability score is heavily influenced by the formula used. Thus, the same set of passages could be identified as being appropriate for 2^{nd} and 5^{th} grade students, based solely upon readability formula selection. The one caveat to these findings is that readability formulas appear to rank a set of passages relatively equally, regardless of the type of formula used.

Progress Monitoring and CST-ELA Scores

Progress monitoring data were collected weekly over a span of six weeks. In order to determine what, if any, extra variance in a high-stakes statewide assessment could be accounted for beyond the winter ORF screening conducted at the schools, the slope from the six weeks of progress monitoring were added to regression equations. The

English Language Arts section of the CSTs, the high-stakes statewide assessment for California, was used in these analyses.

For both 3rd and 5th grades the addition of 6-weeks of ORF progress monitoring did not significantly improve the prediction of CST-ELA scores above the winter ORF screening, regardless of the type of ORF assessment used. There are several explanations for these results. First, the progress monitoring probes were administered to all students in the participating classrooms and student skill was not controlled for in this study. In addition, data on what, if any, additional support or interventions a student was receiving was not available, making interpretation of the effects of these interventions impossible. It is also possible that the school-wide screening data is sufficiently able to predict the CST-ELA scores and that further progress monitoring is redundant, a result that is supported by other research (Schatschneider, Wagner, & Crawford, 2008).

Despite the lack of significance found there were some unexpected results encountered. For 3rd grade students, despite being randomly assigned to one of the three passage type groups (DIBELS Next, AIMSWeb, SDC), there were inconsistencies in the base model when predicting CST-ELA scores. The DIBELS Next group winter ORF Screening scores were significantly related to CST-ELA scores but the AIMSWeb and SDC group scores were not. These results were found before any data from the progress monitoring were included in the equation. Post-hoc exploration found no significant differences between groups on winter ORF Screening scores, $F(2, 43) = .453$, $p = .64$, or CST-ELA scores, $F(2, 43) = .024$, $p = .976$.

The 5th grade results were somewhat more expected in that each group's winter ORF Screening score was significantly related to its CST-ELA outcome before the addition of the progress monitoring. When the slope from 6-weeks of progress monitoring was added there was not a significant increase in variance explained. The small ΔR^2 observed for the passage sets (DIBELS Next = .067; AIMSWeb = .016; SDC = .016) provide little practical or statistical gain over the screening data alone. Further post-hoc analyses revealed other unexpected trends in the data. The average slope for 3rd grade students was .83 (SD = 2.20) indicating average gains of just under a word per week for all 3rd grade students. This finding is similar to previous estimations of 1 word per week growth for 3rd grade students (Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993). However, for the 5th grade students the average slope was -.65 (SD = 2.48) indicating a loss of over a half of a word per week across all students; a result that contradicts the expected growth of 0.5 words per week (Fuchs, et al.). Further post-hoc analyses reveal that there was a significant difference between reading passage groups' slope for 5th grade, $F(2,60) = 5.75$, $p = .005$. Follow-up analyses, using Tukey HSD, revealed that there was a significant difference between the mean slope for the DIBELS Next group (M = .742) and the AIMSWeb (M = -1.42) and SDC (M = -1.27). There was not a significant difference between the AIMSWeb and SDC slopes. Why the AIMSWeb and SDC groups showed negative growth over the course of the six weeks is unclear. One possible explanation is test fatigue. Though each student was assessed for approximately three minutes per week, the week-after-week assessments may have led to decreased motivation to read as well as possible. Regardless of the cause, there is not a clear

explanation as to why this trend would be observed for the AIMSWeb and SDC groups and not the DIBELS Next group since passage type was randomly assigned.

Frequency of Progress Monitoring Data

In order to contribute to the literature on the frequency and number of progress monitoring data points (e.g. Jenkins, 2009), several of these variables were evaluated in this study. Following the recommendations of Agresti and Finlay (1997), regression models were run to compare two reduced models (bi-weekly and first three weeks) to a complete model (all six weeks). The progress monitoring data were derived from the median score of three ORF reading passages, the standard technique used in ORF research and practice. Then, in order to examine the stability of the median-of-three-probes approach the previous analyses were run using only one, randomly selected, ORF score from each week. For all equations, the outcome measure used was the CST-ELA.

Median-of-Three Approach. Using the median of three ORF scores, the complete and reduced regression models were analyzed for each grade and for each passage type. For the 3rd grade students, none of the three passage sets were significant predictors of the CST-ELA when using the complete model (i.e. all six data points) and only one (DIBELS Next) was significant for 5th grade students. Thus, six weeks of weekly ORF progress monitoring provided little information to predict CST-ELA outcomes, with the exception of DIBELS Next in 5th grade. The finding that progress monitoring data are not predictor of later reading ability is analogous to findings from other research (e.g. Schatschneider, Wagner, & Crawford, 2008).

The reduced models (biweekly and first-three-weeks-only) were found to be more parsimonious alternatives to the complete models for all of the 5th grade passage sets. In addition, in 3rd grade, the first-three-weeks-only model was a more parsimonious alternative to the DIBELS Next complete model. The change in degrees of freedom, and the subsequent change in the F-statistic, is a possible explanation for why the reduced models were generally more successful in predicting CST-ELA.

One-Randomly-Selected-Probe Approach. The previous analyses were replicated using one randomly selected data point from each progress monitoring week instead of the median of three. The results indicate areas of similarity and discrepancy between the two approaches. The unexpected results from the 3rd grade slope regression models were replicated. Again, the winter ORF screening scores were significant predictors of CST-ELA for the DIBELS Next group but not for the AIMSWeb or SDC groups. The addition of the progress monitoring slope did not significantly improve the model for any of the groups. For the 5th grade, the results using the one-randomly-selected-probe approach replicated the results from the median-of-three approach. Thus, there were no observable differences in outcome when using the one-randomly-selected-probe approach as compared to the median-of-three approach. However, the unexpected lack of relationship between the screening data and CST-ELA outcomes for two of the groups makes interpretation of these results difficult.

When the one-randomly-selected-probe procedure was applied to the complete and reduced (bi-weekly and first-three-weeks-only) models both similarities and differences between this approach and the median-of-three approach were observed.

Examination of the complete models reveals that there were more significant predictors of CST-ELA when using the one-randomly-selected-probe approach. Using this approach, one 3rd grade (AIMSWeb) and all 5th grade sets were significant – an increase from the one complete model that was significant (3rd grade DIBELS Next) using the median-of-three approach. Examination of the reduced models showed similarities based upon grade. For the 3rd grade group, one of the reduced models was significant in both the median (DIBELS Next first-three-weeks) and one-probe (AIMSWeb biweekly). All other reduced models were non-significant in 3rd grade. In 5th grade, however, all the reduced models were significant for all passage sets for both the median and one-probe approaches.

There were some observed differences between the median-of-three and one-randomly-selected-probe approaches. Specifically, there was a change in variance accounted for by the complete models. For the 3rd grade passages, the amount of variance explained by the complete models decreased for the DIBELS Next (83% v. 37.1%) and the SDC (37.8% v. 27.7%). However, for the AIMSWeb passages there was an increase in variance explained (77.3% v. 95.1%). The same pattern was observed with the 5th grade passages. The amount of variance explained decreased for the DIBELS Next (67.9% v. 66.8) and SDC (78.9% v. 69.6%) and increased for the AIMSWeb (64.8% v. 77.5%). These reduction in variance explained as observed in the DIBELS Next and SDC sets is expected due to the expected increase in variability of weekly scores using only one probe. Unlike the relatively stable decrease observed in the SDC passages, there were large differences between the DIBELS Next 3rd and 5th grade changes in

variance (45.9% and 1.1%, respectively). This discrepancy could be due to variations in passage difficulty included in the DIBELS Next weekly sets. The DIBELS Next ORF passage sets are comprised of three passages that have been labeled as having easy, middle and hard levels of difficulty (Good & Kaminski, 2010). Thus, it is possible that there was more stability in the type of passage that was randomly selected in the 5th grade set as opposed to the 3rd grade set. Another possibility is that there is a greater level of equivalence for the 5th grade passages. Why the AIMSWeb passages showed an increase in variance explained when the number of probes was reduced is unclear.

Differences by Grade. Though not a focus of this research there were some anecdotal differences observed between the 3rd and 5th grade outcomes, regardless of the number of probes used. The relationship between the ORF progress monitoring sets appear to be different based upon grade. For 3rd grade, there was a general lack of significant relationship between the models and the CST-ELAs. Out of a total of 18 models (one complete and two reduced for each passage set, using both probe selection approaches), there were 15 non-significant models. The three exceptions, as previously described, were: DIBELS Next – first-three-weeks using the median approach; AIMSWeb complete and biweekly using the one-probe approach. In 5th grade, the opposite trend is found. Of the 18 models, 16 were significant with the AIMSWeb and SDC complete models using the median approach being the only non-significant models. These results indicate some sort of discrepancy between the 5th grade and 3rd grade ORF passage sets ability to predict CST-ELA when they're used as progress monitoring tools.

This discrepancy needs to be more thoroughly investigated to determine if these results can be replicated using alternative outcome measures and across different grades.

Limitations

There were several limitations that must be considered when evaluating these results. First, the data come from two schools in one school district in Southern California. The student population in this district has a relatively high percentage of both English Language Learners and students who received free and reduced lunches. Thus, the population used in this research may not be representative of all populations. Second, the number of students available was limited. This fact, coupled with the research design which separated students by grade and then by passage type resulted in a relatively small number of students reading a given passage type in a given grade. Future research should attempt to increase the number of students used and thereby increase the power of the results found. A final limitation involves some statistical assumptions that were violated. Specifically, in research questions three and four, the use of weekly ORF probes as individual predictors in the regression equations led to high levels of multicollinearity. Given the nature of the data collected and the design of the research question the multicollinearity is both expected and unavoidable. When interpreting the results this violation should be considered. For research question one, there was a violation of the Kendall's statistic. The probes administered during the second phase were not counterbalanced. In order to verify the current results, future research should replicate using a counterbalanced presentation of the probes.

Conclusions

This research examined the equivalency and predictive utility of ORF progress monitoring probes. The results indicate that there is a clear discrepancy in the identification of equivalent passages. The readability formulas, more often than not, provided similar rankings of passage difficulty, however, their mean grade-level scores varied greatly and were often 3-5 grade levels different than the identified passage level. In contrast, the G-theory results showed that the large majority of the variability observed was attributable to students and not to the passages. These results, coupled with other research (Poncy, et al, 2005; Christ & Ardoin, 2009), lend support to the idea that statistical equating of ORF passages should be considered as a more reliable and informative option than traditional readability formulas when evaluating passage equivalency.

There are less firm conclusions regarding the use of ORF progress monitoring probes to predict the CSTs, a high-stakes statewide assessment. The addition of six weeks of progress monitoring did not significantly improve the ability to predict CST-ELA scores above winter ORF screening data. The results seem to lend support to previous research (Schatschneider, Wagner, & Crawford, 2008) indicating that later reading achievement can be accurately measured with one-time assessments. The difference between the relationship between progress monitoring and CST-ELAs for 3rd and 5th grade passages, regardless of the number of probes selected, requires further investigation.

References

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). Upper Saddle River, NJ: Prentice Hall

Allen, M. J., & Yen, W. M. (2001*). Introduction to measurement theory.* Long Grove, IL: Waveland Press.

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

Ardoin, S. P., & Christ, T. J. (2009). Curriculum-based measurement of oral reading: Standard errors associated with progress monitoring outcomes from DIBELS, AIMSweb, and an experimental passage set. *School Psychology Review, 38*(2), 266-283.

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

Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulas: a critique. *Language & Communication, 21,* 285-301

Barger, J. (2003). *Comparing the DIBELS oral reading fluency indicator and the North Carolina end of grade reading assessment* (Technical Report). Asheville, NC: North Carolina Teacher Academy.

Betts, J., Pickart, M., & Heistad, D. (2009). An investigation of the psychometric

evidence of CBM-R passage equivalence: Utility of readability statistics and equating for alternate forms. *Journal of School Psychology, 47,* 1-17.

Bruce, B., & Rubin, B. (1988) Readability formulas: Matching tool and task. In Davison, A. and Green, G.M. (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered* (pp. 5-22). Lawrence Erlbaum, Hillsdale, NJ.

Buck, J., & Torgesen, J. (2003). *The relationship between performance on a measure of Oral reading fluency and performance on the Florida Comprehensive Assessment Test* (FCRR Technical Report No. 1). Tallahassee, FL: Florida Center of Reading Research.

Burns, M. K., Appleton, J. J., & Stehouwer, J. D. (2005). Meta-analytic review of responsiveness-to-intervention research: Examining field-based and research-implemented models. *Journal of Psychoeducational Assessment, 23,* 381-394.

California Board of Education, Standards and Assessment Division. (2009, March). *California Standards Tests: Technical Report: Spring 2008 Administration.* Retrieved April 12, 2008 from the California Department of Education Web site: http://www.cde.ca.gov/ta/tg/sr/documents/csttechrpt08.pdf

Christ, T. J. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimating standard error of the slope to construct confidence intervals. *School Psychology Review, 35*(1), 128-133.

Christ, T. J., & Ardoin, S. P. (2009). Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology, 47,* 55-75.

Christ, T. J., & Silberglitt, B. (2007). Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency. *School Psychology Review, 36*(1), 130-146.

Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment, 7*(4), 303-323.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability scores and profiles. NewYork: Wiley.

Dale, E. & Chall, J. (1948). A formula for predicting readability. *Educational Research Bulletin, 27,* 37-54.

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

Deno, S. L., Fuchs, L. S., Marston, D., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30*(4), 507-524.

Flesch, R. F. (1948). A new readability yardstick. *Journal of Applied Psychology, 32,* 221-233.

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46,* 315-342.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M.

(1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curve analysis. *Journal of Educational Psychology, 88*(1), 3-17.

Fry, E. (1977). Fry's readability graph: Clarifications, validity, and extensions to level 17. *Journal of Reading, 21,* 242-252.

Fry, E. (2002). Readability versus leveling. *The Reading Teacher, 56*(3), 286-291.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*(3), 199-208.

Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research & Practice, 13*(4), 204-219.

Fuchs, L. S., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment. *The School Psychology Review, 28*(4), 659-671.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (2007). Using curriculum-based measurement to inform reading instruction. *Reading and Writing, 20*(6), 553-567.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22,* 27-48.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239-256.

Fuchs, D., Fuchs, L. S., McMaster, K. N., & Al Otaiba, S. (2003). Identifying children at risk for reading failure: Curriculum-based measurement and the dual-discrepancy approach. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 431-449). New York: Guilford Press.

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

Gersten, R., Compton, D., Connor, C.M., Dimino, J., Santoro, L., Linan-Thompson, S., & Tilly, W.D. (2008). *Assisting students struggling with reading: Response to Intervention and multi-tier intervention for reading in the primary grades. A practice guide.* (NCEE 2009-4045). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved March 26th, 2010 from http://ies.ed.gov/ncee/wwc/publications/practiceguides/.

Good, R. H., & Kaminski, R. A. (1996). Assessment for instructional decisions: Toward a proactive/prevention model of decision-making for early literacy skills. *School Psychology Quarterly, 11*, 326-336.

Good, R. H. & Kaminski, R. A. (2002). DIBELS oral reading fluency passages for first through third grades (Technical Report No. 10). Eugene, OR: University of Oregon.

Gresham, F. (2007). Evolution of the response-to -intervention concept: Empirical

foundations and recent developments. In S. Jimerson, M. Burns, & A. VanDerHeyden (Eds.), Handbook of response to intervention: The science and practice of assessment and intervention (pp. 10-24). New York: Springer.

Gunning, R. (1968). *The technique of clear writing.* New York: McGraw-Hill.

Hintze, J. M. (2005). Psychometircs of Direct Observation. *School Psychology Review, 34*(4), 507-519.

Hintze, J. M., & Christ, T. J. (2004). An examination of variability as a function of passage variance in CBM progress monitoring. *School Psychology Review, 33,* 204-217.

Hintze, J. M., Owen, S. V., Shapiro, E. S., & Daly, E. J. (2000). Generalizability of oral reading fluency measures: Application of G theory to curriculum-based measurement. *School Psychology Quarterly, 15*(1), 52-68.

Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review, 34*(3), 372-386.

Hosp, M. K., & Fuchs, L. S., (2005). Using CBM as an indicator of decoding, word reading and comprehension: Do the relations change with grade? *School Psychology Review, 34*(1), 9-26.

Howe, K. B., & Shinn, M. M. (2002). Standard Reading Assessment Passages (RAPs) for use in general outcome measurement: A manual describing development and technical features. Retrieved from http://www.aimsweb.com/uploads/pdfs/passagestechnicalmanual.pdf

Jenkins, J. R., Graff, J. J., & Miglioretti, D. L. (2009). Estimating reading growth using intermittent CBM progress monitoring. *Exceptional Children, 75*(2), 151-163.

Jenkins, J. R., Zumeta, R., Dupree, O., & Johnson, K. (2005). Measuring gains in reading ability with passage reading fluency. *Learning Disabilities Research & Practice, 20*(4), 245-253.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80*, 437-447.

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

Marcoulides, G. A. 1999: Generalizability theory: picking up where the Rasch IRT model leaves off? In Embretson, S.E. & Hershberger, S.L. (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 129-52). Mahwah, NJ: Lawrence Erlbaum.

Marcoulides, G. A. (2000). Generalizability theory. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling (*pp. 527-551). San Diego: Academic Press.

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

McGlinchey, M. T., & Hixson, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33*(2), 193-203.

McLaughlin, G. (1969). SMOG grading: A new readability formula. *Journal of Reading, 22,* 639-646.

Micro Power & Light Co., 2000 (2000). Readability Calculations. Dallas, TX

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, D.C.: National Institute of Child Health and Human Development.

Olinghouse, N. G., Lambert, W., & Compton, D. L. (2006). Monitoring children with reading disabilities' response to phonics intervention: Are there differences between intervention aligned and general skill progress monitoring assessments? *Exceptional Children, 73*(1), 90-106.

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

Perfetti, C. A. (1992). The representation problem in reading acquisition. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 145-174). Hillsdale, NJ: Erlbaum.

Poncy, B. C., Skinner, C., H., Axtell, P. K. (2005). An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum-based measurement. *Journal of Psychoeducational Assessment, 23,* 326-338.

Powers, R. D., Sumner, W. A., & Kearl, B. E. (1958). A recalculation of four readability formulas. *Journal of Educational Psychology, 49,* 99-105.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seldenberg, M. S. (2002).

How should reading be taught? *Scientific American, 286*(3), 84.

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*(6), 427-469.

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

Shaw, R. & Shaw, D. (2002) *DIBELS oral reading fluency-based indicators of third grade reading skills for Colorado State Assessment Program (CSAP)* (Technical Report). Eugene, OR: University of Oregon.

Sheskin, D. (2007). *Handbook of parametric and nonparametric statistical procedures* (4^{th} ed.) Boca Raton, FL: Chapman & Hall.

Shinn, M. R. (1989). Identifying and defining academic problems: CBM screening and eligibility procedures. In: Shinn MR, editor Curriculum-based measurement: assessing special children. New York: Guilford Press; 1989. p. 90-129.

Shinn, M. R. (1998). *Advanced applications of curriculum-based measurement.* New York: Guilford Press.

Shinn, M. R. (2002). Best practices in using curriculum-based measurement in a problem-solving model. In A. Thomas, & J. Grimes (Eds.), Best practices in school psychology (pp. 671−698). Bethesda, MD: National Association of School Psychologists.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). Boston, MA: McGraw-Hill.

Silberglitt, B., Burns, M. K., Madyun, N. H., & Lail, K. E. (2006). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. *Psychology in the Schools, 43*(5), 527-535.

Spache, G. (1953). A new readability formula for primary-grade reading materials. *The Elementary School Journal, 53*(7), 410-413.

Speece, D. L., & Ritchey, K. D. (2005). A longitudinal study of the development of oral reading fluency in young children at risk for reading failure. *Journal of Learning Disabilities, 38*(5), 387-399.

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

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360-406.

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*(8), 795-819.

Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of School Psychology, 40*(1), 7-26.

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K., &

Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34*(1), 35-58.

Vander Meer, C. D., Lentz, F. E., & Stollar, S. (2005). *The relationship between oral reading Fluency and Ohio proficiency testing in reading* (Technical Report). Eugene, OR: University of Oregon.

Vellutino, F. R., Scanlon, D. M., & Lyon, G. R. (2000). Differentiating between difficult-to-remediate and readily remediated poor readers. *Journal of Learning Disabilities,33*(3), 223 -238

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

Vellutino, F. R., Scanlon, D. M., & Tanzman, M. S. (1998). The case for early intervention in diagnosing specific reading disability. *Journal of School Psychology*, *36*(4), 367-398.

Vellutino, F. R, Scanlon, D., & Zhang, H. (2007). Identifying reading disability based on response to intervention: Evidence from early intervention research. In S. Jimerson, M. Burns, & A. VanDer-Heyden (Eds.), Handbook of response to intervention: The science and practice of assessment and intervention (pp. 185-211). New York: Springer.

Vellutino, F. R., Scanlon, D. M., Zhang, H., & Schatschneider, C. (2008). Using response to kindergarten and first grade intervention to identify children at-risk for long-term reading difficulties. *Reading and Writing*, *21*(4), 437-480.

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

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). Tempe, AZ: Tempe School District No. 3.

Table 1

Descriptive Statistics for CST-ELA and Winter ORF Screening

Group	*N*	*M(SD)*	Skewness	Kurtosis
CST-ELA				
Third Grade	47	326.6 (49.0)	-.17	-.36
Fifth Grade	63	318 (45.3)	.07	-.69
Winter ORF Screening				
Third Grade	47	79.6 (41.2)	.11	-.97
Fifth Grade	64	107.6 (37.9)	-.04	-.69

Note. CST-ELA = California Standards Test – English Language Arts. Winter ORF Screening data derived from AIMSWeb district-wide screening collected at the schools.

Table 2

Descriptive Statistics for Readability Formulas for 3rd Grade Passages

	DIBELS Next[a]			AIMSWeb[b]			SDC[b]		
Formula	*M (SD)*	*Min*	*Max*	*M (SD)*	*Min*	*Max*	*M (SD)*	*Min*	*Max*
Flesch	2.68 (.43)	1.9	3.4	1.70 (.50)	0.7	2.8	4.33 (1.68)	2.1	7.6
FOG	4.75 (.43)	4.0	5.7	3.73 (.48)	3.1	4.8	6.61 (1.65)	4.4	10.4
Powers	4.32 (.17)	4.0	4.6	4.00 (.19)	3.6	4.4	4.62 (.49)	3.9	5.4
SMOG	6.39 (.32)	5.8	7.2	5.45 (.49)	4.8	6.5	7.01 (1.36)	5.0	9.2
FORCAST	8.76 (.53)	7.6	9.7	8.25 (.51)	7.2	9.4	8.49 (.83)	7.1	9.8
Spache	3.79 (.25)	3.1	4.6	3.20 (.33)	2.5	3.8	3.07 (.69)	2.2	5.2
Fry	3.77 (.81)	2.2	5.3	1.90 (.55)	0.8	3.0	4.14 (1.81)	1.2	6.8

Note. SDC = School District Curriculum Passages.
[a]Data derived from 21 passages.
[b]Data derived from 20 passages.

Table 3

Descriptive Statistics for Readability Formulas for 5th Grade Passages

	DIBELS Next[a]			AIMSWeb[b]			SDC[b]		
Formula	*M (SD)*	*Min*	*Max*	*M (SD)*	*Min*	*Max*	*M (SD)*	*Min*	*Max*
Flesch	3.80 (.39)	3.3	4.6	2.65 (.72)	1.5	4.4	7.05 (2.24)	2.9	11.9
FOG	5.85 (.60)	4.9	6.9	4.87 (.94)	3.6	6.8	9.57 (2.32)	5.6	14.2
Powers	4.72 (.16)	4.5	5.0	4.30 (.26)	3.9	5.0	5.32 (.61)	4.1	6.2
SMOG	7.24 (.42)	6.4	8.0	6.33 (.73)	5.2	7.7	9.15 (1.66)	6.1	11.3
FORCAST	9.36 (.45)	8.7	10.3	8.58 (.56)	7.7	9.9	8.98 (.82)	7.5	10.2
Dale-Chall	5.31 (.66)	4.5	7.5	3.84 (.69)	2.7	5.7	5.75 (1.55)	2.9	9.1
Fry	5.36 (.53)	4.5	6.3	3.34 (1.06)	1.6	6.0	7.00 (2.47)	2.8	14.0

Note. SDC = School District Curriculum Passages.
[a]Data derived from 21 passages.
[b]Data derived from 20 passages.

Table 4

Kendall's Coefficient of Concordance by Grade and Passage Type

Group	*W*	χ^2	df	*p*
DIBELS Next				
3rd Grade	.610	85.35	20	<.001
5th Grade	.413	58.87	20	<.001
AIMSWeb				
3rd Grade	.568	75.60	19	<.001
5th Grade	.635	84.40	19	<.001
School District Curriculum				
3rd Grade	.797	105.94	19	<.001
5th Grade	.836	111.23	19	<.001

Table 5

Wilcoxon Pairwise Comparisons for DIBELS Next 3rd Grade Passages

Readability Formula	1	2	3	4	5	6	7
1. Flesch	-	-2.37	-1.87	-2.58	-.22	-.85	-.02
2. FOG		-	-2.81	-.20	-2.67	-.89	-3.30*
3. Powers			-	-3.14	-.97	-.56	-2.11
4. SMOG				-	-2.70	-.85	-3.62*
5. Forcast					-	-1.25	-.16
6. Spache						-	-1.05
7. Fry							-

Note. Significance level is set at * $p < .001$ to account for multiple comparisons.

Table 6

Wilcoxon Pairwise Comparisons for DIBELS Next 5th Grade Passages

Readability Formula	1	2	3	4	5	6	7
1. Flesch	-	-3.28*	-2.62	-3.66*	-3.24*	-3.51*	-3.82*
2. FOG		-	-3.30*	-2.77	-1.14	-2.40	-1.60
3. Powers			-	-3.86*	-3.63*	-3.53*	-4.00*
4. SMOG				-	-1.52	-2.25	-2.23
5. Forcast					-	-2.87	-.11
6. DaleChall						-	-3.14
7. Fry							-

Note. Significance level is set at * $p < .001$ to account for multiple comparisons.

Table 7

Wilcoxon Pairwise Comparisons for AIMSWeb 5th Grade Passages

Readability Formula	1	2	3	4	5	6	7
1. Flesch	-	-3.26*	-2.70	-3.57*	-3.12	-2.46	-4.00*
2. FOG		-	-.34	-2.40	-1.14	-3.40*	-2.34
3. Powers			-	-1.30	-1.99	-3.43*	-2.42
4. SMOG				-	-.66	-3.32*	-1.84
5. Forcast					-	-3.70*	-1.37
6. DaleChall						-	-3.83*
7. Fry							-

Note. Significance level is set at * $p < .001$ to account for multiple comparisons.

Table 8

Wilcoxon Pairwise Comparisons for School District Curriculum 3rd Grade Passages

Readability Formula	1	2	3	4	5	6	7
1. Flesch	-	-.869	-2.93	-2.94	-2.67	-2.36	-2.58
2. FOG		-	-2.23	-2.80	-2.22	-2.54	-2.31
3. Powers			-	-.18	-2.05	-2.05	-.13
4. SMOG				-	-1.96	-1.96	-.80
5. Forcast					-	-.18	-2.13
6. Spache						-	-2.13
7. Fry							-

Note. Significance level is set at * $p < .001$ to account for multiple comparisons.

Table 9

Wilcoxon Pairwise Comparisons for School District Curriculum 5th Grade Passages

Readability Formula	1	2	3	4	5	6	7
1. Flesch	-	-2.17	-2.69	-2.87	-3.04	-3.18*	-2.48
2. FOG		-	-3.11	-3.11	-3.11	-3.18*	-2.94
3. Powers			-	-2.87	-2.83	-2.83	-.80
4. SMOG				-	-2.34	-1.71	-2.97
5. Forcast					-	-.25	-2.97
6. DaleChall						-	-2.70
7. Fry							-

Note. Significance level is set at * $p < .001$ to account for multiple comparisons.

Table 10

Predictors of California State Test-English Language Arts Using DIBELS Next

	Model 1			Model 2		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*
			Third Grade			
Winter ORF Screening	.795	.321	.581*	1.02	.33	.745*
Progress Monitoring Slope[a]				21.14	13.21	.390
R^2		.338			.463	
F		6.13*			4.74*	
ΔR^2					.125	
			Fifth Grade			
Winter ORF Screening	.788	.193	.693**	.841	.189	.739**
Progress Monitoring Slope[a]				5.09	3.20	.263
R^2		.480			.547	
F		16.60**			10.27**	
ΔR^2					.067	

Note. Third grade $n = 14$. Fifth grade $n = 20$.
[a]ORF data are calculated using the median-of-three approach.
* $p < .05$; ** $p < .01$

Table 11

Predictors of California State Test-English Language Arts Using AIMSWeb

	Model 1			Model 2		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*
		Third Grade				
Winter ORF Screening	.266	.244	.289	.208	.245	.226*
Progress Monitoring Slope[a]				5.65	4.73	.318
R^2		.084			.181	
F		1.19			1.33	
ΔR^2					.097	
		Fifth Grade				
Winter ORF Screening	.862	.169	.760**	.834	.174	.735**
Progress Monitoring Slope[a]				-2.06	2.46	-.128
R^2		.577			.593	
F		25.91**			13.10**	
ΔR^2					.016	

Note. Third grade $n = 15$. Fifth grade $n = 21$.
[a]ORF data are calculated using the median-of-three approach.
* $p < .05$; ** $p < .01$

Table 12

Predictors of California State Test-English Language Arts Using SDC

	Model 1			Model 2		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*
		Third Grade				
Winter ORF Screening	.-282	.378	-.203	-.295	.368	-.212
Progress Monitoring Slope[a]				-4.72	3.58	-.348
R^2		.041			.162	
F		.556			1.61	
ΔR^2					.121	
		Fifth Grade				
Winter ORF Screening	.730	.225	.597**	.814	.260	.666**
Progress Monitoring Slope[a]				-3.84	5.60	.145
R^2		.356			.372	
F		10.50**			5.34*	
ΔR^2					.016	

Note. Predictors for CST-ELA using School District Curriculum Passages. Third grade $n = 15$. Fifth grade $n = 21$.
[a]ORF data are calculated using the median-of-three approach.
* $p < .05$; ** $p < .01$

Table 13

Predictors of CST-ELA Using Complete and Reduced Median-of-Three Progress Monitoring Models: DIBELS Next - 3rd Grade

	Complete Model			Bi-weekly			First Three Weeks		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Week 1[a]	-9.24	2.94	-6.00*				-6.46	2.34	-4.20*
Week 2[a]	.672	2.99	.395	1.75	3.72	1.03	2.22	2.78	1.31
Week 3[a]	4.84	4.40	3.29				5.07	2.73	3.45
Week 4[a]	-.01	5.12	-.004	-1.77	3.49	-1.11			
Week 5[a]	3.09	2.04	1.81						
Week 6[a]	1.86	2.19	1.13	1.09	3.24	.66			
R^2		.83			.36			.67	
F		3.92			1.51			5.30*	
ΔR^2					-.46			-.16	

Note. CST-ELA = California State Test-English Language Arts. All models have $n = 12$.
[a]Progress monitoring data are based on median of three passages
* $p < .05$; ** $p < .01$

Table 14

Predictors of CST-ELA Using Complete and Reduced Median-of-Three Progress Monitoring Models: DIBELS Next - 5th Grade

	Complete Model			Bi-weekly			First Three Weeks		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Week 1[a]	.043	.802	.038				-.233	.761	-.206
Week 2[a]	-1.0	1.45	-.772	-.359	1.24	-.28	-.686	1.16	-.529
Week 3[a]	.917	1.36	.681				1.93	.884	1.43*
Week 4[a]	-.24	1.12	-.184	1.05	.968	.802			
Week 5[a]	1.22	.687	.951						
Week 6[a]	.055	1.39	.046	.206	1.44	.172			
R^2		.679			.502			.575	
F		3.52*			4.37*			5.85**	
ΔR^2					-.177			-.104	

Note. CST-ELA = California State Test-English Language Arts. All models have $n = 17$.
[a]Progress monitoring data are based on median of three passages
* $p < .05$; ** $p < .01$

Table 15

Predictors of CST-ELA Using Complete and Reduced Median-of-Three Progress Monitoring Models: AIMSWeb – 3rd Grade

	Complete Model			Bi-weekly			First Three Weeks		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Week 1[a]	-1.49	1.75	-1.34				.076	1.19	.068
Week 2[a]	-2.86	1.82	-2.75	-3.22	1.90	-3.10	-1.44	1.69	-1.38
Week 3[a]	-1.79	2.16	-1.65				1.91	1.95	1.76
Week 4[a]	.941	2.37	.874	2.23	1.44	2.07			
Week 5[a]	3.78	1.65	3.77						
Week 6[a]	1.55	1.84	1.59	1.35	1.52	1.09			
R^2		.773			.468			.297	
F		2.27			2.05			.985	
ΔR^2					-.305			-.476	

Note. CST-ELA = California State Test-English. All models have $n = 17$.
[a]Progress monitoring data are based on median of three passages.
* $p < .05$; ** $p < .01$

Table 16

Predictors of CST-ELA Using Complete and Reduced Median-of-Three Progress Monitoring Models: AIMSWeb – 5th Grade

	Complete Model			Bi-weekly			First Three Weeks		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Week 1[a]	.854	1.01	.705				1.31	.838	1.08
Week 2[a]	.122	.76	.102	.665	.586	.558	.045	.682	.038
Week 3[a]	.490	1.14	.422				-.455	.799	-.392
Week 4[a]	-.592	1.04	-.496	-.848	.911	-.711			
Week 5[a]	-1.02	.932	-.825						
Week 6[a]	.934	.832	.810	1.04	.696	.897			
R^2		.648			.551			.565	
F		2.76			4.90*			5.20*	
ΔR^2					-.097			-.083	

Note. CST-ELA = California State Test-English. All models have $n = 16$.
[a]Progress monitoring data are based on median of three passages.
* $p < .05$; ** $p < .01$

Table 17

Predictors of CST-ELA Using Complete and Reduced Median-of-Three Progress Monitoring Models: School District Curriculum – 3rd Grade

	Complete Model			Bi-weekly			First Three Weeks		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Week 1[a]	1.33	1.13	1.09				1.66	1.07	1.35
Week 2[a]	.084	1.54	.07	.480	.870	.39	-1.40	1.14	-1.14
Week 3[a]	1.50	1.65	1.28				-.102	.838	-.086
Week 4[a]	1.21	1.33	.84	.961	1.24	.671			
Week 5[a]	-2.18	1.86	-1.79						
Week 6[a]	-1.56	1.37	1.34	-1.14	1.35	-.97			
R^2		.378			.066			.191	
F		.809			.260			.866	
ΔR^2					-.311			-.187	

Note. CST-ELA = California State Test-English. All models have $n = 15$.
[a]Progress monitoring data are based on median of three passages.
* $p < .05$; ** $p < .01$

Table 18

Predictors of CST-ELA Using Complete and Reduced Median-of-Three Progress Monitoring Models: School District Curriculum – 5th Grade

	Complete Model			Bi-weekly			First Three Weeks		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Week 1[a]	-1.24	1.53	-1.09				-.487	1.26	-.427
Week 2[a]	-.954	1.57	-.831	-1.58	1.42	-1.38	.690	1.24	-.602
Week 3[a]	.419	1.41	.342				.598	.932	.488
Week 4[a]	2.48	1.28	2.19	2.05	1.10	1.81			
Week 5[a]	-.675	1.37	-.506						
Week 6[a]	.728	2.29	.575	.324	1.18	.256			
R^2		.623			.553			.442	
F		2.76			5.36*			.3.44*	
ΔR^2					-.07			-.181	

Note. CST-ELA = California State Test-English. All models have $n = 17$.
[a]Progress monitoring data are based on median of three passages.
* $p < .05$; ** $p < .01$

Table 19

Predictors of California State Test-English Language Arts Using DIBELS Next

	Model 1			Model 2		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*
			Third Grade			
Winter ORF Screening	.795	.321	.581*	.790	.339	.578*
Progress Monitoring Slope[a]				-.780	7.37	-.026
R^2		.338*			.339	
F		6.13*			2.82	
ΔR^2					.001	
			Fifth Grade			
Winter ORF Screening	.788	.193	.693**	.820	.204	.721**
Progress Monitoring Slope[a]				1.39	2.31	.108
R^2		.480**			.491	
F		16.61**			8.19**	
ΔR^2					.011	

Note. Third grade $n = 14$. Fifth grade $n = 20$.
[a]ORF data are calculated using the one-randomly-chosen-probe approach.
* $p < .05$; ** $p < .01$

Table 20

Predictors of California State Test-English Language Arts Using AIMSWeb

	Model 1			Model 2		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*
		Third Grade				
Winter ORF Screening	.266	.244	.289	.297	.226	.323
Progress Monitoring Slope[a]				5.09	2.80	.447
R^2		.084			.282	
F		1.19			2.36	
ΔR^2					.198	
		Fifth Grade				
Winter ORF Screening	.862	.169	.760**	.865	.169	.762**
Progress Monitoring Slope[a]				-1.88	1.88	-.150
R^2		.577			.599	
F		25.91**			13.47**	
ΔR^2					.022	

Note. Third grade $n = 15$. Fifth grade $n = 21$.
[a]ORF data are calculated using the one-randomly-chosen-probe approach.
* $p < .05$; ** $p < .01$

Table 21

Predictors of California State Test-English Language Arts Using SDC

	Model 1			Model 2		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*
		Third Grade				
Winter ORF Screening	.-282	.378	-.203	-.405	.378	-.291
Progress Monitoring Slope[a]				-3.93	2.91	-.367
R^2		.041			.168	
F		.556			1.21	
ΔR^2					.127	
		Fifth Grade				
Winter ORF Screening	.730	.225	.597**	.814	.260	.666**
Progress Monitoring Slope[a]				-3.84	5.60	.145
R^2		.356			.356	
F		10.50**			4.98*	
ΔR^2					.000	

Note. Predictors for CST-ELA using School District Curriculum Passages. Third grade $n = 15$. Fifth grade $n = 21$.
[a]ORF data are calculated using the one-randomly-selected-probe approach.
* $p < .05$; ** $p < .01$

Table 22

Predictors of CST-ELA Using Complete and Reduced One-Randomly-Selected-Passage Progress Monitoring Models: DIBELS Next – 3rd Grade

	Complete Model			Bi-weekly			First Three Weeks		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Week 1[a]	-.930	3.24	-.588				-.200	2.14	-.126
Week 2[a]	.676	4.84	.382	.377	2.75	.213	1.44	3.62	.815
Week 3[a]	-.629	5.88	-.425				-.166	3.56	-.112
Week 4[a]	1.33	7.62	.803	-.262	3.17	-.158			
Week 5[a]	.395	2.78	.243						
Week 6[a]	.299	4.77	.184	.885	2.05	.546			
R^2		.371			.357			.340	
F		.491			1.48			1.38	
ΔR^2					-.01			-.03	

Note. CST-ELA = California State Test-English Language Arts. All models have $n = 12$.
[a]Progress monitoring data are based on one randomly selected passage
* $p < .05$; ** $p < .01$

Table 23

Predictors of CST-ELA Using Complete and Reduced One-Randomly-Selected-Passage Progress Monitoring Model: DIBELS Next – 5th Grade

	Complete Model			Bi-weekly			First Three Weeks		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Week 1[a]	.391	.594	.350				-.149	.588	-.134
Week 2[a]	.505	.687	.410	.485	.667	.394	.751	.709	.610
Week 3[a]	-.071	.931	-.062				.238	.643	.207
Week 4[a]	.057	.604	.045	.510	.607	.404			
Week 5[a]	1.43	.634	1.03*						
Week 6[a]	-1.10	1.11	-.937	-.091	.692	-.078			
R^2		.668			.493			.468	
F		3.36*			4.22*			3.81*	
ΔR^2					-.175			-.20	

Note. CST-ELA = California State Test-English Language Arts. All models have $n = 17$.
[a]Progress monitoring data are based on one randomly selected passage
* $p < .05$; ** $p < .01$

Table 24

Predictors of CST-ELA Using Complete and Reduced One-Randomly-Selected-Passage Progress Monitoring Models: AIMSWeb – 3rd Grade

	Complete Model			Bi-weekly			First Three Weeks		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Week 1[a]	1.46	.542	1.33				1.83	1.35	1.67
Week 2[a]	-3.14	.627	-3.20*	-2.55	.555	-2.59**	-1.06	1.21	-1.08
Week 3[a]	-1.08	.671	-.922				-.145	1.26	-.124
Week 4[a]	1.17	.920	1.15	1.86	.949	1.83			
Week 5[a]	.467	.853	.446						
Week 6[a]	1.68	.652	1.74	1.25	.773	1.30			
R^2		.951			.849			.359	
F		12.81*			13.13**			1.31	
ΔR^2					-.101			-.592*	

Note. CST-ELA = California State Test-English Language Arts. All models have $n = 11$.
[a]Progress monitoring data are based on one randomly selected passage
* $p < .05$; ** $p < .01$

Table 25

Predictors of CST-ELA Using Complete and Reduced One-Randomly-Selected-Passage Progress Monitoring Models: AIMSWeb – 5th Grade

	Complete Model			Bi-weekly			First Three Weeks		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Week 1[a]	1.26	.514	1.19*				.978	.420	.926*
Week 2[a]	-.432	.588	-.354	.764	.599	.626	-.092	.551	-.076
Week 3[a]	.520	.556	.472				-.082	.460	-.075
Week 4[a]	.512	.566	.481	-.230	.535	-.216			
Week 5[a]	-1.24	.617	-1.04						
Week 6[a]	-.036	.482	-.033	.323	.421	.299			
R^2		.775			.485			.636	
F		5.17*			3.76*			6.99**	
ΔR^2					-.290			-.139	

Note. CST-ELA = California State Test-English Language Arts. All models have $n = 16$.
[a]Progress monitoring data are based on one randomly selected passage
* $p < .05$; ** $p < .01$

Table 26

Predictors of CST-ELA Using Complete and Reduced One-Randomly-Selected-Passage Progress Monitoring Models: School District Curriculum – 3rd Grade

	Complete Model			Bi-weekly			First Three Weeks		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Week 1[a]	1.04	.871	.980				.762	.657	.719
Week 2[a]	.749	.894	.682	.514	.590	.469	.816	.770	.744
Week 3[a]	-1.33	1.41	-1.25				-1.42	.860	-1.35
Week 4[a]	.819	1.56	.642	-.159	1.05	-.124			
Week 5[a]	-.371	.976	-.382						
Week 6[a]	-1.04	1.28	-.630	-.552	1.05	-.336			
R^2		.277			.097			.218	
F		.511			.396			1.02	
ΔR^2					-.180			-.060	

Note. CST-ELA = California State Test-English Language Arts. All models have $n = 15$.
[a]Progress monitoring data are based on one randomly selected passage
* $p < .05$; ** $p < .01$

Table 27

Predictors of CST-ELA Using Complete and Reduced One-Randomly-Selected-Passage Progress Monitoring Models: School District Curriculum – 5th Grade

	Complete Model			Bi-weekly			First Three Weeks		
Variable	*B*	*SE B*	*β*	*B*	*SE B*	*β*	*B*	*SE B*	*β*
Week 1[a]	-.862	1.04	-.738				-.269	.562	-.230
Week 2[a]	-.038	.851	-.030	.908	.713	.730	.098	.603	.079
Week 3[a]	1.19	.501	1.03*				1.09	.370	.942*
Week 4[a]	.204	.609	.197	-.352	.614	-.340			
Week 5[a]	.094	.761	.083						
Week 6[a]	.357	.471	.308	.360	.426	.311			
R^2		.696			.477			.669	
F		3.82*			.3.96*			8.76**	
ΔR^2					-.219			-.027	

Note. CST-ELA = California State Test-English Language Arts. All models have $n = 17$.
[a]Progress monitoring data are based on one randomly selected passage
* $p < .05$; ** $p < .01$