UNIVERSITY OF CALIFORNIA
RIVERSIDE

Screening English Learners with Oral Reading Fluency: The Prevalence of Word Callers

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

Doctor of Philosophy

in

Education

by

Kerri Theresa Colleen Knight-Teague

December 2011

Dissertation Committee:
Dr. Mike Vanderwood, Chairperson
Dr. Rollanda O'Connor
Dr. Sara Castro-Olivo

The Dissertation of Kerri Theresa Colleen Knight-Teague is approved:

_____

_____

_____
Committee Chairperson

University of California, Riverside

Dedication

This work is dedicated to my amazing husband, Rob Teague, for being so patient and supporting me throughout graduate school. It is accurate to say that much of this process was possible because of you. Thank you so much.

Thank you to my advisor, Dr. Mike Vanderwood, for the feedback and practical advice you have given not only on this project, but also throughout my graduate school career. It is truly appreciated.

Additionally, thank you to Dr. Rollanda O'Connor, for giving thorough and insightful feedback on my work. You have challenged me to become a better researcher and writer as a result.

ABSTRACT OF THE DISSERTATION

Screening English Learners with Oral Reading Fluency:
The Prevalence of Word Callers

by

Kerri Theresa Colleen Knight-Teague

Doctor of Philosophy, Graduate Program in Education
University of California, Riverside, December 2011
Dr. Mike Vanderwood, Chairperson

Oral reading fluency (ORF) as an indicator of reading comprehension was examined for a sample of third and fifth grade English learners (ELs). The impact of English language proficiency on the relationship between ORF and reading comprehension, and the prevalence of word callers, or students who are fluent readers, but do not comprehend at a proportionate level, were examined using a series of regression and predictive accuracy analyses. Additionally, teacher judgments of participants' reading skills were explored with a focus on the accuracy of teachers' word caller nominations in their classrooms. Results showed that word callers emerged, though infrequently, and that there were inaccuracies associated with teachers' judgments of reading skills.

Table of Contents

List of Tables

List of Figures

Screening English Learners with Oral Reading Fluency:

The Prevalence of Word Callers

A primary goal of reading connected text is to comprehend, or create meaning from the words on the page. Good readers read text accurately and with adequate pace to promote comprehension (Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003). For struggling readers, precise allocation of instructional resources is needed to facilitate this type of reading achievement and to prevent a course of failure that is likely to occur without intervention (Juel, 1988).

National assessments have suggested that many ELs are struggling readers. For example, the National Assessment of Educational Progress (NAEP) in 2007 showed that 71% of fourth grade ELs were performing in the "below basic" category in reading in contrast to 30% of native English speakers (NESs) that fell into the "below basic" category (U.S. Department of Education, IES, NCES, NAEP, 2009). Additionally, the population of ELs in the nation, and in California particularly is quite large. According to national reports from the years 1996 to 2006, the population of students that are learning English in United States schools (i.e., pre-kindergarten through 12th grade) has increased by about 57% (National Clearinghouse for English Language Acquisition, 2007). Over half of the nation's English learners (ELs) were attending schools in the West as of the year 2000 (U.S. Department of Education, Institute of Education Sciences [IES], National Center for Education Statistics, 2004) and about 24% of students (1,513,233 out of 6,252,031) enrolled in California schools were ELs during the 2008-2009 school year (California Department of Education [CDE], 2009a).

There are potential solutions to help improve reading achievement, including that of ELs. Several studies document the effectiveness of providing reading intervention to NESs early to improve reading skills (e.g., Torgesen, 2002; Wanzek & Vaughn, 2007) and in the later grades to remediate deficits that were not addressed early or emerged later (Torgesen, Alexander, Wagner, Rashotte,Voeller, & Conway, 2001). Improvement in reading-related skills has been observed in samples of ELs when students were provided with evidence-based reading interventions similar to those that have been used with NESs (Lesaux & Siegel, 2003; Linan-Thompson, Vaughn, Prater, & Cirino, 2006; Lovett, et al., 2008; Vaughn, et al., 2006). Screening for deficits in reading skills to allocate intervention services is a data-based decision making process that is necessary in today's educational climate (Good, Simmons, & Kame'enui, 2001) and can help solve problems of underachievement due to lack of appropriately matched instruction by assessing all students in a particular grade level and identifying those that need intervention the most (Gersten et al., 2008). However, it is crucial that screening tools measure skills that are indicators of reading achievement (Deno, 2003) and accurately classify most students, including ELs.

Current screening recommendations for ELs parallel those made for NESs (i.e., Gersten, Baker, Shanahan, Linan-Thompson, Collins, & Scarcella, 2007) and include screening for reading problems with English measures of phonological processing, letter knowledge, word reading, and text reading when instruction is given in English (Gersten et al.). These recommendations are aligned with studies that have shown several early literacy skills measured in English are indicative of later English reading-related

outcomes for ELs (e.g., Fien, Baker, Smolkowski, Smith, Kame'enui, & Beck, 2008; Gottardo, Collins, Baciu, & Gebotys, 2008; Gottardo & Mueller, 2009; Quiroga et al., 2002; Vanderwood, Linklater, & Healy, 2008). However, these studies have focused on younger ELs (i.e., kindergarten through third grade) and early literacy measures (e.g., phonological awareness, letter knowledge, word reading). Fewer studies examine appropriate screening tools for older ELs.

Oral reading fluency (ORF) is one assessment that has been investigated as a screening tool for a wide range of elementary school aged students (e.g., first through sixth grade; Reschly, Busch, Betts, Deno & Long, 2009). ORF is considered a curriculum-based measurement (CBM). CBM is based on the idea of measuring a global goal, or entire skill set from the beginning of the instructional process (Fuchs, 2004). Because of this conceptualization, CBM has been termed a type of "general outcome measurement" (Fuchs & Deno, 1991). CBM is technically sound, standardized, is comprised of direct observations, is time efficient, and uses multiple equivalent samples (Deno, 2003). Because CBM is intended to be short in duration and the focus of assessment is decidedly not comprehensive, CBM is termed an "indicator" of overall performance (Shinn & Bamonto, 1998).

ORF has received a great amount of attention in the literature due to its association with reading comprehension (Fuchs, Fuchs, Hosp, & Jenkins, 2001). Though administration procedures vary, there is consensus between two groups (i.e., Good & Kaminski, 2002; Shinn & Shinn, 2002) that ORF is a timed task in which a student reads from a passage of connected, meaningful text for a discrete time period (usually one

minute). The examiner records the number of words read correctly in this time period. Omissions, mispronunciations, substitutions, and hesitations longer than three seconds are usually counted as incorrect. Researchers label the task many ways (e.g., curriculum-based measurement of reading, reading fluency, connected text fluency, etc.). For continuity, the term ORF is used here to describe all measures that time the reading of connected text and generate a score representing words read correct scaled by time. ORF is not meant to be confused with single word reading tasks that are sometimes labeled as "fluency" measures (e.g., Proctor, Carlo, August, and Snow; 2005) since single word reading tasks function differently (Jenkins et al., 2003).

The use of ORF is framed within reading theories that assert automatic and fluent decoding skills are a core component of reading comprehension (Gough, Hoover, & Peterson, 1996; Hoover & Gough, 1990) and facilitate higher order comprehension skills (e.g., LaBerge & Samuels, 1974; Nathan & Stanovich, 1991; Posner & Snyder, 1975). Theories of automaticity and efficiency (e.g., Perfetti, 1985; Posner & Snyder, 1975) differ in regard to the process by which higher order activities interact with word recognition, but they are similar because they assert fluent word recognition allows comprehension to take place more effectively (Fuchs et al., 2001). Based on the theoretical context, many studies have investigated the validity of ORF scores for the applied purpose of screening within a tiered model of intervention support.

The Validity of ORF Scores

Guidelines for Evaluating Validity

Deno (2003) suggested that ORF and other CBMs can fulfill a wide variety of purposes (e.g., screening, progress monitoring). When the purpose of an assessment is to screen for risk status, the focus of evaluation is different than other tests. For example, less concern is directed at the differentiation between groups across a broad range of levels and more attention is given to dichotomizing groups into "at risk" or "not at risk" status (Rathvon, 2004). However, the procedures by which validity evidence is accumulated are not different than any other assessment.

Glover and Albers (2007) recommended evaluating a screening tool's appropriateness for intended use, adequacy of technical characteristics, and usability. Technical adequacy includes the investigation of both reliability and validity. By including aspects of appropriateness and usability, these recommendations highlight that there is more to the evaluation of a screening tool than its technical characteristics alone. This notion is aligned with the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/ACME], 1999) guidelines for establishing validity for all assessments. The guidelines in the test standards are based on Messick's (1994; 1995) definition of validity. Messick defined validity as both the meaning and the use of a test score. Thus, a score is valid if its interpretation is commensurate with what the test is intended to measure (score meaning) and the

presence of the test score results in beneficial consequences or action taken on behalf of the examinee (score use; e.g., access to better matched instruction).

Score meaning. Several issues are critical when the meaning of a screening score is considered in a study. One is the population for which the screening tool will be used. For ELs, limited English language proficiency has the potential to create construct irrelevant variance, or error due to a factor that is unrelated to the construct measured by the test (Messick, 1995). This irrelevant variance (in this case, limited English language proficiency) can change the meaning of the test score. For example, the test score might reflect limitations in English language knowledge rather than a measure of the test's targeted construct. For this reason, the test standards indicate that the linguistic background of an examinee must be taken into consideration and validity evidence should be collected for separate linguistic groups when differences in score meaning are suspected (AERA/APA/ACME, 1999) To fulfill this requirement, validity evidence for ORF scores must be accumulated for samples of ELs.

A second issue is the selection of the outcome criterion to which a screening tool will be compared. The outcome should be meaningful, relevant to the purpose of screening, and its psychometric qualities should be adequate (Jenkins et al., 2007). Jenkins and colleagues also suggested that differentiation between state-level tests and published norm-referenced tests (PNRTs; that are nationally normed) as outcomes should be reported because the former are related to standards that are state specific which limits the generalization of the results.

After the selection of an appropriate sample and outcome criterion, a third issue is the method used to establish score meaning. Technical adequacy is the key to the interpretation of a score's meaning. Jenkins, Hudson, and Johnson (2007) differentiated between two approaches to accumulate this type of validity evidence for a screening tool: correlation analysis (which authors termed criterion validity) and predictive accuracy. Both approaches are similar in that they evaluate the relationship between a screening tool and an outcome criterion, but differ by the method used to describe the relationship.

The correlation approach uses correlation coefficients and regression analyses to model the strength of the relationship between a screening tool and an outcome criterion, which establishes a meaningful link between the screening tool and target skills for intervention. Jenkins and colleagues acknowledged that correlation studies are important evidence for validity, but are not sufficient on their own because they lack information on the accuracy of the screening tool. Predictive accuracy evaluates the degree to which accurate performance on the outcome criterion was predicted by the screening tool.

Predictive accuracy is considered foundational in screening research and should always be addressed (Jenkins et al., 2007). Information on predictive accuracy is obtained by examining cutoff scores. A cutoff score is "used to classify an examinee as being at risk for failure as defined by the criterion measure" (Rathvon, 2004, p. 20). Ideally, a cutoff score describes a student's present level of performance and indicates whether it is likely they will subsequently do well or poorly on another criterion.

Score use. Glover and Albers (2007) suggested that the following areas be addressed when the use of screening scores are considered: cost effectiveness, feasibility

of use, whether scores are acceptable to multiple stakeholders (e.g., teachers, principals), the infrastructure of the school, accommodations that can be made for students, and whether scores are helpful in guiding treatment (i.e., instructional) decision-making. The methods that might be used for evaluating these constructs were not described.

Using the aforementioned considerations for score meaning and score use as a rubric, most studies that examine the validity of ORF scores document criterion evidence through correlation and regression based analyses. The degree to which studies utilize predictive accuracy, study the applicability of a screening tool to diverse populations such as ELs, and examine score use varies.

ORF Scores and Academic Outcomes

The literature on ORF has included two areas. First, criterion evidence for validity (in both a predictive and concurrent fashion) has been reported to address the meaning of scores. Second, the beneficial consequences of ORF scores and those from other CBMs have been reported to address the use of scores. Studies related to beneficial consequences that occur as a result of using ORF have examined changes in academic performance after the provision of scores. These studies include other CBMs in addition to ORF (e.g., maze, math CBM) and have found the use of CBMs to monitor progress has positive effects on student achievement (Fuchs & Fuchs, 1986; Fuchs, Fuchs, & Hamlett, 1989a; Fuchs, Deno, & Mirkin, 1984; Stecker, Fuchs, & Fuchs, 2005) and teacher behavior (Fuchs, Fuchs, & Hamlett, 1989b). Most studies have focused on the meaning of ORF scores, rather than their use, as a contribution to validity evidence for the purpose of screening and are subsequently described.

Reviews and meta-analyses. Two studies that synthesized the association between ORF and other reading outcomes have been published recently: one examined CBM broadly, yet included ORF (Wayman, Wallace, Wiley, Ticha, & Espin, 2007) and one examined ORF specifically (Reschly et al., 2009). Wayman et al. (2007) conducted a literature synthesis on curriculum-based measurement (CBM) as a whole. This included word identification, ORF, and maze. Maze tasks are another type of CBM and comprise a passage of connected text in which the first sentence is left intact and subsequently every seventh word is deleted and replaced with three possible choices, one of which is the correct answer (Parker, Hasbrouck, & Tindal, 1992). The authors sought to conduct an updated review on the topic since only one other work of large magnitude was published prior to their publication (e.g., Marston, 1989). The authors concluded that the literature they reviewed supported a link between ORF and comprehension at the group level. Additionally, Wayman et al. suggested that maze might be a better measure of progress for older students (i.e., secondary school level) because growth rates were more stable across the literature than those for ORF. Both maze and ORF were comparable in the primary grades in terms of their association with reading outcomes.

Wayman et al. (2007) performed a comprehensive review of the literature on ORF, but did not conduct a quantitative synthesis of the information. Reschly and colleagues (2009) conducted a meta-analysis on ORF as it relates to reading achievement. The authors collected journal articles and technical reports that studied ORF related to some other reading outcome in first through sixth grade. Articles that did not use standardized procedures for the administration of ORF, used achievement scores to

predict ORF, combined data across grades, or were reported in a dissertation or conference proceedings were excluded from the analysis. Reschly and colleagues coded whether the outcome test was state-specific or a PNRT, whether the test was individually or group administered, the type of criterion score (i.e., comprehension, vocabulary, word identification, decoding, or total reading), the amount of time between the administration of ORF and the outcome, and the grade level of the participants.

Reschly and colleagues' (2009) analysis used 289 correlation coefficients; the median coefficient was .68, with an average weighted coefficient (taking study sample size into account) of .67. The authors used a hierarchical linear model to analyze correlation coefficients nested within individual studies. ORF was a significant predictor of both state-level tests ($r = .65$) and PNRTs ($r = .74$). However, the PNRT correlation coefficient was significantly higher than the state-level coefficient. The comparison between individual- and group-administered tests (which only applied to PNRTs since state-level tests were only administered in groups) showed correlation coefficients of .83 and .71, respectively; coefficients were significantly different. Reschly et al. speculated this difference emerged because the magnitude of reliability found in individually administered assessments is greater. Results by grade showed no significant differences among correlation coefficients for first through sixth grade students. Length of time between the administration of ORF and the outcome moderated the magnitude of the relationship between ORF and the outcome: a negative relationship indicated that as time increased correlation coefficients deteriorated. Finally, ORF was found to be a statistically significant predictor of comprehension outcomes as well as those measuring

vocabulary and decoding. However, the relationship between ORF and word identification resulted in a stronger correlation that was significantly different than all others. The authors concluded that ORF and word identification tasks might involve more common processes than the other tasks.

Reschly et al. (2009) concluded that overall results regarding ORF were positive, especially considering its low cost in terms of both time and resources. Results confirmed the relationship between ORF and reading related outcomes across grades. However, no differences in the magnitude of the relationship were found between grades, which conflicts with results reported from individual samples (e.g., Shinn, Good, Knutson, & Tilly, 1992; Wayman et al., 2007).

Single studies. The Wayman et al. (2007) and Reschly et al. (2009) studies synthesized most current research on ORF. However, examining individual ORF validity studies is valuable to examine predictive accuracy (of students into groups based on global reading outcomes) and to examine studies separated into general categories by grade. Reschly et al. noted that researchers have sought to quantify the relationship between ORF and state-level, high-stakes assessments that are most often group-administered in recent years instead of looking at PNRTs of reading. Because of this shift, most presented here focus on state-level assessment as the outcome criterion for screening.

Studies using participants in the primary grades (i.e., first through third) are common. For example, Crawford, Tindal, and Stieber (2001) compared the ORF scores of students during second grade to reading performance on another ORF administration

and the Oregon statewide reading test during their third grade year. The authors used passages derived from students' curriculum that were not equated; the median score from three passages was taken. The coefficients between second grade ORF and the statewide reading test and third grade ORF were .66 and .84, respectively. Second grade ORF was also compared to math performance on the statewide test during third grade and yielded a coefficient of .53. This is limited evidence of divergence (i.e., ORF is more closely associated to other measures of reading than other skills).

In another study, Good et al. (2001) examined a sample of young students in kindergarten through third grade. This study added predictive accuracy to correlation analysis. ORF was used in this study among other early literacy indicators to examine performance on subsequent administrations of ORF as well as the Oregon Statewide Assessment- Reading/Literature (OSA). The authors reported a correlation of .82 between spring administrations of ORF from first and second grades. During the spring of third grade, a coefficient of .67 was reported between ORF and the OSA. These were measured in close proximity. The authors established a cutoff score based on the third grade ORF administration of 191 words. Ninety-six percent of students that met or exceeded the cutoff score had satisfactory or above performance on the OSA.

Goffreda, Diperna, and Pedersen (2009) also included predictive accuracy information in their study by using logistic regression and predictive accuracy indices. These methods allow for predictive accuracy analysis. The authors examined the predictive validity of the early literacy subtests of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002). Four DIBELS subtests were

administered to 67 first grade students and compared to the TerraNova California Achievement Test (CAT) and the Pennsylvania System of School Assessment (PSSA). The following DIBELS subtests were used: Letter Naming Fluency, Phoneme Segmentation Fluency, Nonsense Word Fluency, and ORF. Results indicated that winter first grade ORF scores were significantly correlated with the CAT (administered in second grade) and the PSSA (administered in third grade). Additionally, logistic regression analyses showed that ORF was the only subtest in which risk classification status (i.e., low risk, some risk, high risk) significantly predicted dichotomous performance outcomes (i.e., proficient and not proficient) on outcome measures. Finally, the cutoff scores provided by DIBELS for ORF attained sensitivity and specificity levels (80% and 87%, respectively) that were deemed reasonable by the authors.

Johnson, Jenkins, Petscher, and Catts (2009) also investigated the accuracy of using the DIBELS measures to screen early for reading problems. Johnson and colleagues' study also used young participants and included considerations of base rates when examining predictive accuracy indices. Participants were 12,055 students followed longitudinally from kindergarten to third grade. The authors were interested in end of the year performance during first grade and examined several of the DIBELS early literacy measures in addition to ORF. ORF was included as a screening measure during first grade. Johnson et al. chose SAT-10 results as the first grade spring outcome measure, since test scores on this measure were most related to later performance on the Florida Comprehensive Assessment Test – Sunshine State Standards (FCAT-SSS).

Sensitivity was set to 90% and corresponding specificity, cutoff scores, and hit rate were examined with regard to SAT-10 scores dichotomized at the 40th percentile as the outcome. With sensitivity set at 90%, fall first grade ORF had a specificity of 59%, a cutoff score of 18 words read correct per minute, and hit rate of 70%. Base level hit rate was 71% indicating that if all students were assumed to be not at-risk in the absence of any screening measurement, 71% would be correctly classified. Optimal levels of sensitivity and specificity calculated by logistic regression were 52% and 87%, respectively. Johnson and colleagues (2009) also examined the same data by dichotomizing the SAT-10 outcome at the 20th percentile to attempt to classify very poor readers. With sensitivity set at 90%, specificity was 65%, the cutoff score was 14 words read correct per minute, and hit rate was 67% (compared to 90% predictive accuracy for base rate). Optimal levels of sensitivity and specificity were 11% and 99%, respectively. Considering scores on the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 1997) did not improve accuracy markedly.

Johnson et al. (2009) conducted analyses for ELs and students receiving free or reduced price lunches as separate subgroups. Predictive accuracy revealed lower cutoff scores for ELs and students receiving free or reduced price lunches on ORF than those reported for the larger sample. The authors recommended that schools examine screening results by subgroup. Johnson and colleagues concluded that ORF was a moderate predictor of end of the year first grade performance.

Baker et al. (2008) used another sophisticated method of examining the relationship between ORF and both the Oregon Statewide Reading Assessment (OSRA)

and The Stanford Achievement Test- Tenth Edition (SAT-10) for first through third grade students. Growth curve models were established based on fall, winter, and spring administrations of ORF. Performance on ORF was related to both the OSRA and SAT-10 with correlations between .58 and .82. The results of Baker and colleague's analysis suggested that examining the slope of ORF scores helped improved prediction.

In another similar study, Hintze and Silberglitt (2005) evaluated ORF as a predictor of performance on the Minnesota Comprehensive Assessment (MCA) with a sample of first through third grade students. However, this study used three different data analytic techniques: receiver operator characteristic (ROC) characteristic curves, logistic regression, and discriminant analysis. Part of the purpose was to determine which analysis was best. ORF passages were reported as first through third grade material (Shinn & Shinn, 2002). Students were assessed with ORF a total of eight times: winter and spring of first grade and then fall, winter, and spring for both second and third grade. The MCA was administered during the end of the students' third grade year. First, correlation coefficients were examined between each time point and the MCA. Coefficients ranged from .49 to .69, and generally increased with each time period. The discriminant analysis showed that using each subsequent administration of ORF as the dependent variable resulted in cutoff scores that were incrementally higher. Sensitivity was between 82% and 95%. Using MCA performance as the criterion resulted in a fluctuation of cutoff scores that did not follow a pattern. Additionally, sensitivity was only 65% at its highest using this approach. The results for the logistic regression and ROC curve analysis were similar. The authors also found that cutoff scores generally

resulted in higher levels of specificity than sensitivity. Hintze and Silberglitt concluded that ORF is a good predictor of performance on subsequent measures of ORF and high-stakes assessment, but that subsequent administrations of ORF seem to be a better criterion for establishing cutoff scores rather than using state test scores that are not within close temporal proximity to the ORF administration.

Another group of studies have included samples of primarily third grade students and above. Silberglitt, Burns, Madyun, and Lail (2006) conducted a correlation analysis across a wide range of grade levels. ORF and maze were examined in comparison to the Minnesota Comprehensive Assessment-Reading (MCA-R) and the Basic Standards Test-Reading (BST-R). Students in third, fifth, seventh, and eighth grade were included in the sample. Coefficients between ORF and the outcomes fell between .51 and .71; the coefficients for third and fifth graders were significantly greater than those for seventh and eighth graders. Coefficients between maze and the outcomes fell between .49 and .54.

Shapiro, Keller, Lutz, Santoro, & Hintze (2006) included predictive accuracy in their study. The authors examined ORF passages from AIMSweb (Shinn & Shinn, 2002) in conjunction with the Pennsylvania System of School Assessment (PSSA) with third and fifth grade students. The authors used correlation and predictive accuracy analyses. ORF was measured in the fall, winter, and spring. Correlations between ORF and the PSSA were in the range of .62 to .68; one coefficient between fall ORF and the PSSA was low (.24). Predictive accuracy indices yielded sensitivity and specificity in the range of 67% to 86%.

Similarly, Stage and Jacobsen (2001) examined the ORF performance of 173 students in fourth grade compared to the Washington Assessment of Student Learning (WASL). ORF data was collected in fall, winter, and spring; the WASL was administered in the spring. Stage and Jacobsen were interested in the predictive utility of ORF at individual time points (i.e., fall, winter, spring) and the slope of student growth across time points. Students were administered a single passage at each time point that was drawn from their fourth grade curriculum. The authors did not report a numerical estimate of the reliability.

Results of Stage and Jacobsen's hierarchical linear model indicated that both student level and slope for ORF was significantly different than zero; there was a positive trend in growth across the year. The relationship between individual time points and the WASL, and slope and the WASL was explored through correlation analyses; all correlations were significant and moderate in size (i.e., .26-.44). However, multiple regression analyses showed that slope did not explain a statistically significant amount of additional variance in WASL scores when individual time points were included in the regression equation. Sensitivity and specificity for fall ORF scores were 66% and 76%, respectively. The authors reported that sensitivity and specificity for winter and spring scores were within plus or minus 1% of fall scores, indicating that the predictive accuracy of ORF across the year remained about the same.

Stage and Jacobsen's (2001) study was limited because the authors did not provide reliability data for the ORF passages that were used and only a single passage, rather than a median from three passages was used to calculate words read correctly per

minute. However, results suggested that while screening results were indicative of outcomes, growth over time was not significant for this sample.

Wood (2006) conducted a different analysis using the same type of data. Hierarchical linear modeling (HLM) was used to examine the relationship between ORF and the Colorado Student Assessment Program (CSAP) for students in third, fourth, and fifth grade. Correlations between ORF and the CSAP ranged from .67 to .75 by grade. Results indicated that ORF predicted CSAP scores for all grade levels and a reliable increase in ORF scores was noted cross-sectionally as grade increased. The relationship between ORF and CSAP varied significantly at the classroom level. Wood speculated that this might suggest instructional or teacher variables influence this relationship in some way. The time period between ORF and CSAP administration was at least two months; this would allow several instructional factors to create possible confounds. No significant variation was found across grades indicating that the relationship between ORF and CSAP scores was consistent as grade increased. The lower bound of the 95% confidence interval for the ORF score that was associated with proficiency cutoff score on the CSAP was used to examine predictive accuracy indices. Sensitivity was above 85% for all grade levels and specificity was above 58%.

Two separate studies conducted the same type of work (i.e., comparing ORF to state-level assessments) with very large samples. Roehrig, Petscher, Nettles, Hudson, and Torgesen (2008) analyzed data from 16, 539 third graders and reported correlation coefficients ranging from .70 to .71. McGlinchey and Hixson (2004) analyzed data from 1,362 fourth grade students for eight years cross-sectionally and reported correlation

coefficients ranging from .49 to .81 across years. In both studies ORF yielded sensitivity and specificity in that ranged from about 70% to 90%. However, Jenkins et al. (2007) noted in a critique of these studies that the highest levels of predictive accuracy were often observed with the closest administration of ORF in proximity to the state-level assessment. Those that were farther in advance (i.e., fall to spring prediction) still did a reasonably good job of identifying struggling students with sensitivity of 74% (Roehrig et al., 2008).

Finally, technical reports are another category of studies that are worth mentioning within the context of ORF screening literature. Third grade students have been the focus of most technical reports (i.e., Barger, 2003; Buck & Torgesen, 2003; Shaw & Shaw, 2002; Vander Meer, Lentz, & Stollar, 2005; Wilson, 2005) with the exception of one study (i.e., Vander Meer et al.) that also included fourth grade. Correlation coefficients between ORF and state-level reading assessments from these reports ranged from .60 to .93, dependent on the time period of administration and the outcome. Sensitivity and specificity reported also approximated or exceeded 75%. Wilson included a group of ELs as a subsample within his report and obtained a correlation coefficient of .78 for third grade students whose spring ORF scores were compared to the state-level assessment.

Summary and limitations. The studies on the relationship between ORF and other state-level reading assessments provide initial evidence to support its use as a screening tool from first through third grade. The studies that reported predictive accuracy varied in proximity to Rathvon's (2004) recommendation of above .75. Though

less frequent, studies conducted with older students (fourth and fifth grade) have shown moderate correlations. Though it appears that correlations are generally smaller in magnitude, some results are mixed.

One inconsistency in the ORF literature on screening is whether there is a significant decrease in the association between ORF and reading comprehension as students get older. Stage and Jacobsen's (2001) study suggested this decline by showing that while a point estimate of ORF (as is done in screening) was a significant indicator of performance for fourth grade students; growth over time (as is done in progress monitoring) was not. There is a clear decrease in ORF's utility between elementary and middle school students (i.e., fifth and seventh grade; Silberglitt et al., 2006), but there are conflicting interpretations about this decline during the elementary school years. The prevalent notion has been that ORF declines in association (Jenkins & Jewell, 1993) and is no longer an important indicator of reading comprehension (Yovanoff, Duesbery, Alonzo, & Tindal, 2005; Shinn et al., 1992) as students get older. This is often framed within the context of oral language (and its components) being a consistent and increasingly important indicator of reading comprehension as reading develops (Vellutino, Tunmer, Jaccard, & Chen, 2007).

The studies that have presented conflicting results include Wood's (2006) work that reported the relationship between ORF and the state-level outcome did not vary as a function of grade when third, fourth, and fifth grade students were considered. Reschly et al. (2009) also reported a similar result across first through sixth grade. Additionally, other studies not associated with the screening literature have failed to find a change in

the relationship between ORF and reading comprehension as measured by a PNRT of reading comprehension across first through fourth grade (Hosp & Fuchs, 2005). Conflicting results suggest the need for more studies that focus on the relationship between ORF and meaningful reading outcomes for older students (i.e., fourth grade and above) in conjunction with the purposes for which ORF might be used.

One limitation in this body of work is that few attempts have been made to control the effects of reading instruction or intervention. This is a problem because there is usually a time lapse between the screening period and the administration of the reading outcome in order to be able to make generalizations about the practicality of screening (i.e., students need to be screened far enough in advance to be able to implement intervention). This is meant to help educators make meaningful decisions about instructional needs across time periods in the school year, but it also might introduce confounding instructional variables. Wood's (2006) results indicated that classroom or teacher level variables likely had an influence on the relationship between ORF and the state-level assessment in the study. Another way to control for the effects of time and instruction would be to examine concurrent relationships between ORF and reading outcomes since screening data are interpreted and intervention recommendations are made within close temporal proximity to screening periods.

Other potential pitfalls in synthesizing validity evidence from ORF screening studies include the variability of: materials used to measure ORF, standardized procedures, number of passages with which a score is obtained, and outcome criteria. There are clearly issues that need to be addressed in future study.

Although some of the aforementioned studies examined ELs as a subgroup, there are others that consider ELs as a primary part of the sample. This is a critical step for establishing validity evidence for ELs' ORF scores as indicated by the test standards (AERA, APA, NCME, 1999) and is necessitated by the large population of ELs served in United States' schools (Kindler, 2002).

ORF Scores for ELs

On a limited basis, ORF has been investigated for use with English learners (ELs) with promising results. However, the research base for ELs has not accumulated the same amount of information as the literature available on NESs and does not have a body of work on screening per se. Most studies have included Spanish-speaking ELs as participants. This is likely because Spanish speakers comprise about 79% of ELs nationally (Kindler, 2002) and about 84% of ELs in states like California (CDE, 2009a). Because of the focus on Spanish-speaking ELs, generalizations to other groups of ELs are limited unless otherwise noted. As with NESs, studies have been conducted that examine reliability and criterion evidence for validity with other variables like PNRTs and state-level assessment. On a very limited basis (viz., Muyskens, Betts, Lau, & Marston, 2009) predictive accuracy has been examined. There is still a great amount of work to be done in this area.

Baker and Good (1995) collected data on the following with a sample of second grade NESs and Spanish-speaking ELs: ORF, a standardized reading assessment, language proficiency batteries, and teacher ratings of both reading ability and language proficiency. Twenty ORF passages were administered over a 10-week period (two times

per week). The ORF passages were created by taking text from the students' reading curriculum.

Results of the Baker and Good (1995) study showed that point and level estimates of ORF were not significantly different between NES and EL groups. This showed that within the sample, student groups did not differ on their first ORF score or on their mean level of ORF performance over time. Slope estimates as calculated by an ordinary-least-squares regression line showed significant differences in favor of the EL group, who made more progress over time. Baker and Good also examined the reliability of point, level, and slope estimates. All coefficients were comparable across student groups except for those representing point estimates, which showed a significantly higher coefficient for the EL group. Reliability coefficients for point and level estimates ranged from .87 to .99, which surpass the recommended a criterion of .80 (Salvia, Ysseldyke, & Bolt, 2007) as a standard for the reliability of screening tools. Finally, validity coefficients as compared to the Stanford Diagnostic Reading Test (Karlsen & Gardner, 1985) and to teacher ratings of reading ability were comparable across student groups and ranged from .51-.80.

Studies conducted more recently have suggested that ORF is predictive of reading performance on state-level assessment for ELs (e.g., Muyskens et al., 2009; Wiley & Deno, 2005). Wiley and Deno included third and fifth grade students in their sample. Participants included both NESs and ELs. The EL portion of the sample included speakers of the following languages: Hmong, Somali, and Spanish. Participants were obtained by screening the entire third and fifth grade population at the school with the Basic Academic Skill Samples (BASS; Deno, Maruyama, Espin, & Cohen, 1990) one

minute maze passages to identify the lowest 50% of students during the spring and fall. The lowest 50% of the population was subsequently administered ORF every two weeks from November to May. The ORF passages were drawn from the Standard Reading Passages (Children's Educational Services, 1987). Participants read aloud for one minute from three different passages and the median score was recorded. ORF and maze performance were correlated with the Minnesota Comprehensive Assessment (MCA). However, the authors did not specify which ORF and maze scores were used for the analysis (e.g., the initial fall maze screener, the initial fall ORF score, etc.). This limits the generalizability of the results. Additionally, progress monitoring with ORF was mentioned, but no analysis regarding this data was conducted.

Wiley and Deno (2005) reported that all correlations between maze and the MCA, and ORF and the MCA were significant for both ELs and NESs. For ELs, correlations ranged from .52 to .69; for NESs, correlations ranged from .57 to .73. Regression analyses showed that maze explained a significant amount of variance beyond ORF for third and fifth grade NESs, but not for ELs. The additional amount of variance explained by maze was especially pronounced for fifth grade NESs: explained variance increased by about 20% when maze was included in the regression equation. The authors concluded that ORF was a better predictor of MCA performance for ELs than maze and that maze was a promising tool for NESs.

Muyskens et al. (2009) also examined ORF in conjunction with the MCA. Participants were fifth grade EL students from a variety of native language backgrounds: Spanish, Hmong, and Somali. The students were administered ORF passages in the fall

drawn from their basal curriculum. The methods by which the passages were equated were unclear. A median score from three passages was used. The authors used a combination of simple regression, logistic regression, and ROC curve analysis. The results of the simple regression analysis indicated that ORF was a significant predictor of the MCA. Both intercept and slope were significant. The correlation coefficient was .62 which is similar to the association between ORF and state-level assessment for NESs. Logistic regression was used to yield a cutoff score to examine predictive accuracy indices. The authors presented predictive accuracy in a reverse fashion such that sensitivity was an index of the proportion of those that were predicted to pass and did pass the MCA and specificity was an index of the proportion of those that were not predicted to pass and did not pass the MCA. Using the cutoff score of 111 words read correct per minute sensitivity was 44% and specificity was 89%. Muyskens et al. also used ROC curve analysis to examine the area under the curve (AUC). Values closer to one indicate better predictive accuracy. Using the cutoff score of 111, a value of .78 was obtained. The authors did not attempt to maximize senstivity and specificity by examining other cutoff scores obtained from the ROC curve analysis. They concluded that ORF provided a better specificity index (in their case prediction of those who would actually fail the MCA). However, they did not mention that the analysis they used (i.e., logistic regression) maximizes this type of prediction.

In addition to point estimates of performance, literature on the growth of ELs as measured by ORF is also available. Initial evidence has shown mixed results regarding the progress made by ELs as measured by ORF. Graves, Plasencia-Peinado, Deno, &

Johnson (2005) found that progress made by ELs as measured by ORF is comparable to that reported for NESs. Participants in the Graves et al. study were first grade students that were obtained after ORF was administered to all students across nine classrooms. The three lowest achieving and three highest achieving students were selected from each classroom, along with three randomly selected students that were part of the intermediary and read between 20 and 50 words per minute. The selected participants were administered ORF once weekly for six weeks. ORF passages were derived from a standardized set (e.g., Shinn & Bamato, 1998) and scores were obtained from one passage read aloud. Graves et al. reported an average weekly growth rate of 2.75 words across groups. For the lowest, middle, and highest groups average words gained per week were 2.8, 3.6, and 1.8, respectively. The authors asserted that this progress was comparable to the rates reported for NESs in other studies. These results are in contrast to another study's from the intervention literature (e.g., Linan-Thompson, Cirino, and Vaughn, 2007) where most ELs never attained comparable benchmarks to NESs when progress on ORF was considered.

More recently, Al Otaiba, Petscher, Pappamihiel, Williams, Dyrlund, and Connor (2009) investigated the progress ELs make on ORF from second to third grade. In addition to ORF, the Peabody Picture Vocabulary Test (PPVT-III; Dunn & Dunn, 1997) was also administered to all students in the sample. Hierarchical linear modeling (Raudenbush & Bryk, 2002) was used to examine ORF performance for three subgroups broken up according to English language proficiency: Latino students who were fluent in English and had never received language support services, students enrolled in English as

a second language (ESL) services, and students exited from ESL services. Within each subgroup, students were further subdivided into three groups by educational setting: general education, speech or language delayed, and learning disabled. Results showed a quadratic trend over time for all groups: performance initially accelerated and then leveled off. The ORF performance of ESL students was lower across all time points than the other two groups and never surpassed state benchmarks. Performance on the PPVT was on average one standard deviation below national norms for the ESL group. Additionally, students designated as learning disabled showed significantly lower trends in growth than their peers that were speech/language delayed or general education within the same English language proficiency classification. The authors concluded that because of the differences in trend, ORF shows promise to reliably indicate students that are in need of special services. The lower performance of the ELs in the Al Otaiba et al. (2009) study seem to be consistent with those from Linan-Thompson et al. (2007) and Johnson et al., (2009).

Mean differences in ORF performance for ELs. Results from studies that have used ORF with ELs show that ORF is predictive of later reading performance and has comparable reliability to the coefficients obtained using samples of NESs. Additionally, rates of change over time on ORF initially appear to be similar across ELs and native English speakers when first grade students are considered. However, the level of attainment on ORF seems to be lower than NESs (e.g., Al Otaiba, 2009; Johnson et al., 2009; Linan-Thompson et al, 2007). A mean difference between groups (i.e., ELs and NESs) is not sufficient evidence to conclude that ORF is biased (AERA/APA/ACME,

1999), but it is a critical concern if there are differences in predictive validity and uniform cutoff scores on ORF are used for both NESs and ELs to establish risk status and subsequent access to intervention. Additionally, the factors that might contribute to the mean differences between ELs and NESs are important considerations.

Studies that examine bias do not just consider mean differences, but also consider differences in predictive validity as evidenced by significantly different slopes or intercepts when regression analysis is performed. The examination of bias in ORF when used with diverse populations has been examined. Bias due to ethnicity has been addressed and results have been mixed (e.g., Hintze, Callahan, Matthews, Williams, & Tobin, 2002; Pearce & Gayle, 2009; Kranzler, Miller, & Jordan, 1999). Bias due to linguistic differences was addressed by Klein and Jimerson (2005). The authors examined a variety of other factors in addition to language in their study that included: ethnicity, gender, and socioeconomic status (using free- or reduced-cost lunch status as a proxy). ORF passages (created by the second author) were used in comparison to the Stanford Achievement Test- Ninth Edition (SAT-9; Harcourt Brace & Co., 1997). Participants were in first through third grade, came from Hispanic or Caucasian backgrounds, and spoke Spanish or English as their first language. Three cohorts of students from three separate years was used in the analysis. No Caucasian participants spoke Spanish as their home language so groups were as follows: Hispanic/Spanish home language, Hispanic/English home language, and Caucasian/English home language. Results showed mean differences between Hispanic students that spoke English and Spanish at home on ORF scores and SAT-9 scores: Spanish-speaking students scored significantly lower in

both areas. Additionally, Hispanic students that received free or reduced price lunches showed significantly lower means on both ORF and SAT-9 scores than Hispanic students that were not eligible for such programs. These results showed that group differences existed when the sample was split up by first language and socioeconomic status (SES), after controlling for ethnicity.

Significant slope bias was found between groups of Hispanic, Spanish-speaking students and Caucasian, English-speaking students when all grades were included in the analysis. When each grade was considered separately, intercept bias was found between groups separated by ethnicity and first language. Specifically, ORF tended to overpredict later reading performance for Spanish-speaking, Hispanic students. Results for the analysis that included free or reduced price lunch as a proxy for SES were inconclusive. Significant intercept bias was found for each grade-level cross sectionally, but effect sizes were generally small. Additionally, slope estimates for two separate cohorts of first grade students were significant, but again small in magnitude. Klein and Jimerson (2005) asserted that SES appeared to make relatively small and inconsistent contributions to bias in the current sample. The authors concluded that a combination of first language and ethnicity were the factors that influenced the bias in ORF. The mean differences on ORF scores and errors in prediction were considered in conjunction and authors advised that precautions must be taken when using this instrument with diverse populations.

Klein and Jimerson (2005) noted that their study only examined one criterion (the SAT-9) and to further validate their results other criterion measures should be used. It is also worth noting that the ORF passages used in the study were created by the second

author and had limited psychometric documentation. Overall the study suggested that ORF bias in prediction as a function of language needs to be investigated with other outcome measures and that it is likely that ELs will need a different set of benchmarks to evaluate their performance. Findings from Klein and Jimerson's study are relevant to the current investigation because mean differences in ORF combined with the overestimation of later reading performance suggests that ELs in the sample were not only performing lower than NESs, but also could indicate ORF scores were not representative of the same construct for both groups.

Summary and limitations. The results from the literature on ORF and ELs are promising because there appears to be a relationship between English ORF and English reading outcomes. However, there some limitations in the literature. First, Klein and Jimerson (2005) concluded that SES did not influence the relationship between ORF and the SAT-9 to a large degree. However, considering SES is still important when ELs are the focus of assessment. About 68% of ELs in preschool through fifth grade in the United States were below the poverty level in the year 2000; these rates were more than double those of NESs (Capps, Fix, Murray, Ost, Passel, & Herwantoro, 2005). Poverty is associated with several negative cognitive, academic, and socioemotional outcomes for students and is even linked to lower teacher expectations (McLoyd, 1998). Studies have shown that when gaps in achievement between ELs and NESs are considered, SES explains a large portion of the disparity (Kieffer, 2008) and that the negative effects of SES can diminish over time for ELs that are exposed to high-quality, evidence-based instruction (D'Angiulli, Siegel, Maggi, 2004). These points make the interpretation of

mean differences in ORF scores between ELs and NESs complex: differences in English language proficiency cannot be assumed to be the only influential variable. Studies that report mean differences need to be cautious about overinterpreting results if SES was not included as a covariate in the analysis.

Second, studies that address ORF and ELs have not always considered English vocabulary or oral language in conjunction with ORF. Though the consensus in the literature for NESs suggests ORF is an adequate screening tool for early elementary age students, considering measures of vocabulary or oral language as a supplement might be beneficial for ELs given expected deficits in English vocabulary (August, Carlo, Dressler, & Snow, 2005; Al Otaiba et al., 2009; Carlo et al., 2004; Francis & Rivera, 2007; Proctor et al.; Snow & Kim, 2007) and listening comprehension (Proctor et al.). Studies have shown that both variables are related to reading comprehension for ELs (Proctor et al., 2005; Nakamoto, Lindsey & Manis, 2007).

One way to account for limited English proficiency overall (including vocabulary and listening comprehension deficits) is to examine tests that measure proficiency in conjunction with ORF. Prior to No Child Left Behind (NCLB) English language proficiency tests had several limitations: lack of defined construct, problems with applicability to academic performance, and issues with technical adequacy (Abedi, 2004; Albers, Kenyon, Boals, 2009). A severe limitation to some English language proficiency tests' validity is the failure to classify NESs as proficient in English (Pray, 2005); in this situation the meaning of ELs' performance on such a test is questionable. However, English language proficiency assessments that have been developed in response to NCLB

are higher quality (Abedi, 2004; Abedi, 2007; Albers et al., 2009). States were required to develop their own English language proficiency assessment or use a published assessment that met stringent criteria. Abedi (2007) identified the distinction between basic interpersonal communication skills (BICS) and cognitive academic language proficiency (CALP) as a critical component of the post-NCLB ELP assessments. Specifically, states are required to align the test to English Language Development (ELD) content standards across academic content areas. This helps assure that CALP will be assessed in an English language proficiency assessment and results will more readily translate into what an EL is capable of understanding in the classroom environment, rather than informal interactions. Additionally, states are mandated to collect data on progress annually to measure goals towards fluent English language proficiency. Given the availability of English language proficiency data studies that include EL participants are bolstered when this variable is considered.

Criticism of ORF

There is a great deal of literature that has confirmed ORF is a tool that holds promise for screening (Reschly et al., 2009; Wayman et al., 2007). As is the case with any assessment there are limitations. For ORF, there is evidence that suggests a decrease in utility as students get older (Yovanoff, Duesbery, Alonzo, & Tindal, 2005; Shinn et al., 1992). However, some criticism has been directed at the use of ORF regardless of age. This criticism is associated with face validity: the nature of the ORF task does not tap the skills that educators commonly associate with reading comprehension because it only directly measures rate and accuracy of decoding connected text (Shinn & Bamonto,

1998). Both educators (see Foegen, Espin, Allinder, Markell, 2001) and researchers (e.g., Kamii & Manning, 2005; Samuels, 2007) have asserted ORF is not capable of yielding information about a complex construct such as reading comprehension.

The term *word caller* has been used to describe the hypothetical profile of students who "efficiently decode words but do so without comparable comprehension taking place," (Meisinger, Bradley, Schwanenflugel, Kuhn, & Morris, 2009, p. 147-148). Dependent on how the phrase "efficiently decode words" is defined, a student that is a word caller might be inaccurately classified as not at risk for reading problems when he or she is experiencing significant deficits in reading comprehension if screening for intervention only included ORF. Word callers are not exclusively associated with ORF, but with measures of word reading in general. Word callers were initially associated with instruction and were speculated to emerge if there was an unbalanced focus on decoding (Meisinger et al., 2009; Stanovich, 1986). The validity of the unbalanced literacy argument is questionable today since balanced literacy instruction is supported by research (National Reading Panel, 2000) and is encouraged in schools in order to meet state accountability requirements (i.e., No Child Left Behind). Additionally, Nathan and Stanovich (1991) made the point that decoding words quickly and accurately is never a negative attribute (even if it were to result from specific instructional practices) because evidence suggests that word reading automatically activates corresponding word meanings as long as they are already solidified in memory. Thus, fluent reading would only be problematic in the screening context if ORF scores fail to accurately predict reading comprehension outcomes. Recent research on word callers suggests that many

teachers might believe this is the case because they nominate students in their classrooms as word callers when given the opportunity (Hamilton & Shinn, 2003; Meisinger et al., 2009; Meisinger, Bradley, Schwanenflugel, & Kuhn, 2010).

Teachers' propensity to nominate students as word callers is not surprising given the existence of student profiles that show accurate word decoding and poor comprehension occurring together (e.g., Dewitz & Dewitz, 2003; Stothard & Hulme, 1992). No difference criterion was set in these studies: differences were defined by word reading accuracy and comprehension performance that were not identical as measured by instructional level or age equivalents. Meisinger et al. (2009) categorized three factors that might lead to the word caller profile: hyperlexia, severe deficits in linguistic comprehension, and deficits in English language proficiency (i.e., ELs). The first factor, hyperlexia, is associated with autism spectrum disorders. A student with hyperlexia is described as having "exceptional word-reading ability above that expected given their IQ, and at a higher level than their ability to comprehend and integrate words," (Newman, Macomer, Naples, Babitz, Volkmar, & Grigorenko, 2007, p. 760). However, hyperlexia is associated with single word reading (Grigorenko, Klin, Pauls, Senft, Hooper, & Volkmar, 2002) and not with tools like ORF that require decoding in context. Additionally, hyperlexia is very rare and occurs in approximately 5-10% of the population of students with autism spectrum disorders (Burd, Kerbeshian, & Fisher, 1985).

The next two factors Meisinger and colleagues (2009) described as possible correlates to word calling are similar since they involve deficits in linguistic

comprehension. This is an important issue to consider since Nathan and Stanovich (1991) asserted that during reading, word meanings are activated automatically only when they are already solidified in memory. Word meanings might not be solidified in memory for many students.

Meisinger and colleagues' (2009) suggested a deficit in linguistic comprehension as their second factor that might lead to word calling when NESs are considered. Evidence suggests that poor comprehension sometimes occurs in the presence of adequate word reading skills (e.g., Stothard & Hulme, 1992). A deficit in linguistic comprehension has been implicated for students whose reading comprehension is extremely poor, but whose decoding or word-recognition skills are adequate (Catts & Hogan, 2003). This profile has occurred in very low frequencies in observed samples (Shankweiler et al., 1999). Additionally, the inclusion of ORF or similar connected text reading tasks to define this profile is inconsistent.

The third and final factor related to word calling proposed by Meisinger and colleagues (2009) was limited English proficiency, which affects ELs. An argument can be made that when an EL is presented with an ORF passage, he or she might be able to decode words in the text, or read them by sight without knowing their meanings because of a deficit in English vocabulary (e.g., August et al., 2005; Al Otaiba et al., 2009; Carlo et al., 2004; Snow & Kim, 2007). Dependent on the rate at which he or she reads, this profile might represent that of a word caller: a student that can read fluently, but does not comprehend at a proportionate level. However, no study to date has investigated word callers in the EL population.

One factor that Meisinger et al. (2009) did not explicitly mention in relation to word callers was age or grade level of students. Given the studies that suggest the decline in utility of ORF as students age (e.g., Shinn et al., 1992; Jenkins & Jewell, 1993) it is reasonable to hypothesize that word callers might exist in more appreciable numbers within groups of older students. This makes examining the prevalence of word callers an important issue for confirming or disconfirming the validity of ORF scores for specific grade levels.

In order to achieve clarity on word caller issues in samples of NESs, two areas have been addressed, (a) the prevalence of students that meet the profile of a word caller and (b) the prevalence and accuracy of teacher word caller nominations. Since the debate on word callers questions the validity of ORF, these research areas can be linked to interpretations of validity: score meaning and score use (Messick, 1994; Messick, 1995). For example, determining the actual prevalence of word callers in a population is important because appreciable numbers of word callers pose a threat to the meaning of ORF scores. Similarly, investigating the frequency of teacher nominations is critical because nominations of word callers pose a threat to the usability of scores: if teachers do not believe ORF is a valid measure of students' ability, information from ORF is unlikely to be used for instructional decision-making or planning.

Empirical Investigations of Word Callers

Currently, three studies have been published that directly address word callers; all have used participants that were NESs and have not found large numbers of word callers in the early elementary years. However, prevalence increased in the upper grades.

Hamilton and Shinn (2003) sought to investigate word callers by asking third grade teachers to identify students in their classes that met the following definition: "student who can read fluently, but has difficulty comprehending text." The authors' rationale for using teacher nominations was based on the idea that a word caller is a common misconception among teachers, rather than a veritable student profile. In addition to a word caller, each teacher was asked to identify a student that read as fluently as the word caller, but had no problems with comprehension. Pairs of students from each classroom were divided into two groups: word callers and similarly fluent peers. These groups were compared with respect to ORF, a maze task, the Passage Comprehension subtest of the Woodcock Reading Mastery Test (WRMT; Woodcock, 1987), and the Comprehension Oral Question Answering Test (CQT; Jenkins et al., 1986). ORF and maze passages were taken from CQT folktales and standardized directions were given. The median score from three administrations was used for ORF. Passage equivalence for CQT folktales was not discussed in the article, but passages were drawn from a previous study (i.e., Jenkins, Heliotis, Haynes, Stein, & Beck, 1986). The authors hypothesized that if the teacher-identified word callers embodied the operational definition, their fluency scores would be similar to those of fluent peers, but their comprehension scores would be significantly lower than fluent peers.

Results indicated that patterns of performance were significantly lower on all measures for students nominated as word callers. The students nominated as word callers not only performed significantly lower than their fluent peers on measures of reading comprehension, but on ORF as well. Individual cases were also examined within the

study: the authors noted that only a single student within the sample (n = 66) might fit the word caller profile. This student obtained a higher score on ORF and answered one to two items less on each of the comprehension measures than his or her matched similarly fluent peer. Hamilton and Shinn asserted that the argument that this single student "did not comprehend" would be faulty given that his or her Passage Comprehension score fell at the 30th percentile. However, since the authors used a subjective definition for word caller and the group was nominated by teacher selection alone, it is impossible to identify or disconfirm individual word callers with any amount of certainty.

The accuracy of teacher judgments in the Hamilton and Shinn study (2003) was also examined. Teachers were asked to predict how their students would perform on all measures except the Passage Comprehension subtest of the WRMT. On average, teachers significantly overestimated performance for both groups of students (word callers and similarly fluent peers) for all measures considered. The authors explored the possibility that teachers were using the terms *accuracy* and *fluency* interchangeably after the study took place. The sample was comparable in terms of accuracy, but no data were collected on teacher judgments of accuracy beforehand, so this determination could not be made.

The Hamilton and Shinn study (2003) was limited because the authors used the term word caller to describe a group in their study, but did not objectively define this profile. This hindered any conclusions about prevalence that could have been made within the sample. Regardless, generalizations cannot be made about the prevalence of the word calling profile in the population since the entire sample was teacher selected. Conclusions from this study are limited to the accuracy of teacher judgments alone:

teachers tended to overestimate performance for all identified students and nominated students that were poor readers (i.e., low on measures of fluency and comprehension) rather than word callers. This evidence does not disconfirm the notion that word callers might exist in the population nor does it address why teachers might nominate students as word callers from their classrooms.

Meisinger et al. (2009) conducted a similar study on word callers more recently. Meisinger and colleagues used two objective criteria to identify word callers. For criterion 1, standard score cutoffs were identified for both fluency on the Gray Oral Reading Test – Fourth Edition (GORT-4; Wiederholt & Bryant, 2001), set to be a standard score greater than or equal to 95 and reading comprehension on the Wechsler Individual Achievement Test – Reading Comprehension subtest (WIAT-RC; Wechsler, 1992), set to be a standard score less than 85. Criterion 2 was derived from literature on children with autism and hyperlexia; students identified by this criterion would have above average fluency (i.e., a standard score of 110 or higher) and a gap in reading comprehension (i.e., a standard score less than 90). Next, Meisinger et al. included second, third, and fifth grade students in their study to determine whether the prevalence of students fitting the criteria for word calling changes as a function of grade. Whole classes were screened instead of teacher- nominated pairs. This allowed the researchers to not only examine the accuracy of teacher judgments about performance, but also to examine the agreement between research-identified and teacher- nominated word callers. Finally, Meisinger et al. examined teacher opinions and knowledge about fluency and comprehension in depth to attempt better understanding about their judgments.

Results of the Meisinger et al. (2009) study showed that in the primary grades (i.e., second and third grade) the prevalence of students that met the word calling definition was very low (.4 – 2.3%). However, when the authors considered a separate sample of third and fifth grade, a significantly greater percentage of fifth grade students were identified as word callers (9.78%) than third grade students (1.82%) using the criterion 1 definition; no relationship was found for the criterion 2 definition because only one fifth grade student fit the profile from the entire sample. All correlations between reading fluency and comprehension were significant and ranged from .51 to .72 across samples; correlations declined as grade increased.

Meisinger et al. (2009) used three different surveys to obtain teacher judgment data. First, they asked teachers to describe their definitions of "fluency" and "comprehension." Authors were interested in whether teachers would include comprehension as part of their definition of fluency and vice versa. Second, teachers were asked to nominate students from their classes that were word callers. They were told that a word callers are students, "who can read fluently but have difficulty comprehending text," (Meisinger et al., p.152). Finally, teachers rated their students on a Likert-type scale that addressed fluency and comprehension. Results showed that teachers were not accurate in identifying students that met the authors' criterion 1 definition of word callers. Using the combined sample of third and fifth grade students, most teacher-nominated word callers (93.3%) were false positives. Teacher-nominated word callers performed significantly lower than their peers on fluency and reading comprehension measured by both standardized assessments and teacher ratings. Most

teachers in the sample (61.9%) described comprehension as an integral part of their definition of fluency, but did not include fluency as a part of their definition of comprehension. No relationship was found between teachers' definitions of fluency and comprehension and the students that they nominated as word callers.

The third study available on word callers placed much more focus on obtaining teacher definitions of the term word caller and gathering information on teacher assessment and instructional practices (Meisinger et al., 2010). Thirty-one participants who were second grade teachers and their 408 students were included in the study. Students were administered three ORF passages from the DIBELS (Good, Kaminski, Smith, Laimon, & Dill, 2001) and the Gates-MacGintie Reading Test—Fourth Edition (GMRT-4; MacGintie, MacGintie, Maria, Dreyer, & Hughes, 2000). The median score from the ORF passages was used. Teachers were given a survey and asked to define the terms "word caller," "reading fluency," and "reading comprehension." Reading fluency and reading comprehension were coded as "basic" or "expanded." Definitions coded as "expanded" were dependent on whether or not teachers included comprehension processes as a part of fluency and vice versa. Additionally, teachers were asked how they would assess and intervene with a student who was a word caller. Responses were coded as "consistent" or "inconsistent" dependent on the teacher's definition of word caller. For example, if a teacher mentioned only "fluency" as part of the definition of word caller, and also mentioned assessment and intervention strategies that addressed fluency alone, the response would have been coded as "consistent." The authors used reading

assessment textbooks common to teacher preparation programs to guide their judgments about the appropriateness of assessment and intervention.

The results indicated that 1.2% of the second grade students in the sample met the researchers' definitional criteria for a word caller, which was ORF performance at or above the "some risk" DIBELS category (i.e., greater than or equal to 70 words read correctly per minute) combined with a GMRT-4 standard score of 85 or less. Despite the low prevalence, many teachers nominated their students as word callers (i.e., 24.8% of the student sample). There was little association between teacher-nominated and researcher-identified word callers. About 1% of the students nominated by their teachers as word callers had data to confirm such a profile. Subsequent analyses showed no patterns of performance that suggested mean differences in ORF and reading comprehension between teacher-nominated word callers and their peers. Additionally, the relationship between ORF and comprehension in the teacher nominated word caller sample was positive (r = .62).

Most teachers in the sample provided a basic definition of fluency that did not include any mention of comprehension (45.2%), yet some teachers did include comprehension in their definition (38.7%; Meisinger et al., 2010). However, no teachers mentioned fluency in their definitions of comprehension. Teachers' perceptions of the term word caller were overall consistent with the Meisinger et al. (2009) definition. In contrast, about 25.8% of teachers considered word callers to be dysfluent (i.e., low fluency rate) readers and about 22.6% of teachers considered word callers to be poor readers (i.e., low performance on fluency and reading comprehension measures).

Teachers' descriptions of assessment and intervention practices were largely consistent with their definitions of word caller.

Limitations. There are limitations to the studies that addressed word callers. The Hamilton and Shinn (2003) study was limited in that the sample included only teacher nominated word callers. However, both studies by Meisinger and colleagues (2009; 2010) improved upon this and included a comparison of teacher-nominated and research-identified word callers.

Although it is not a limitation per se, future studies should expand the literature base on word callers by including ELs in the sample. Meisinger et al. (2009) suggested that ELs might emerge as word callers due to deficits in English vocabulary, but did not include ELs in their study. Next steps need to include ELs in order to determine whether the prevalence of word callers changes based on the consideration of native language and English language proficiency as recommended by test standards (AERA, APA, NCME, 1999). This is an important issue related to technical adequacy that will confirm or disconfirm additional validity evidence for ELs' ORF scores.

The way teacher judgments about student performance were collected in all of the word caller studies was limited. In the first two studies (i.e., Hamilton & Shinn, 2003; Meisinger et al., 2009) teachers were asked to nominate word callers, but instead identified students that were poor readers (i.e., students that performed below average on both fluency and comprehension measures). However, the studies did not give teachers the option to nominate poor readers. For example, Meisinger et al. (2009) asked teachers to rate both fluency and comprehension for every student, but only asked teachers to list

the names of potential word callers. Hamilton and Shinn (2003) only asked for word caller nominations. Meisinger et al. (2010) only asked teachers to define the term word caller. Although the goal of all three studies was to determine the existence of a word caller population, the limited response sets that were presented to teachers could have constrained their responses. For this reason it is somewhat misleading to report that teachers in all studies were inaccurate in their judgments since teachers did a relatively decent job of identifying students that were struggling. The results might have been different had teachers been given the option to place struggling readers into either a word caller or poor reader category.

There are other issues related to the use of ORF scores (as opposed to the meaning), which are a critical aspect of validity (AERA, APA, NCME, 1999) that have not been fully addressed by the literature available on word callers. Meisinger et al. (2010) quantified teachers' use of "appropriate" assessment strategies given each teacher's judgment about a student's reading fluency and comprehension performance. However, teacher judgments about the usefulness of ORF for all students in a sample (regardless of their teacher-nomination status) have not been quantified. Given that teachers readily nominated students as word callers and most students did not fit the word caller profile (Hamilton & Shinn, 2003; Meisinger et al., 2009; Meisinger et al., 2010) it is pertinent to know whether teachers perceive ORF scores for these students as useful. If scores are not perceived as useful, it is unlikely that teachers will value using ORF as a screening tool for intervention decisions.

Study Purpose

Since large proportions of word callers in a sample would be a threat to the validity of ORF scores, this study was broadly meant to contribute to the validation of ORF scores for ELs. In line with the test standards' (APA, AERA, NCME, 1999) definition of validity, both the meaning and use of ORF scores will be addressed in the research questions.

The first set of questions relate to the meaning of ORF scores for EL participants by addressing the prevalence of word callers and the consideration of English language proficiency as a moderator. Past work that has investigated ORF with EL populations has not included English language proficiency in analyses. In the current study, ELs' English language proficiency was quantified and its relationship with ORF and comprehension was investigated. Since very few studies have investigated word callers and ORF explicitly, and no published studies on word callers have used ELs as participants, a primary purpose was to assess the prevalence of word callers among ELs. Additionally, both third and fifth grade students were included in the sample since there is limited evidence on the validity of ORF with older ELs (making the inclusion of fifth grade relevant). The following questions were addressed using the California Standards Tests-English Language Arts-Reading Comprehension subscale (CST-ELA-RC; Educational Testing Service [ETS] & CDE, 2009) as the reading comprehension outcome. Third and fifth grade data were analyzed separately for each question.

- Research Question 1: Is the relationship between ORF and CST-ELA-RC moderated by English language proficiency?

- Research Question 2a: What is the prevalence of word callers among a sample of ELs for two different definitional criteria?
- Research Question 2b: What is the prevalence of word callers (for two definitional criteria) when levels of English language proficiency are considered?

The next set of questions was meant to address the use of ORF scores by considering teacher judgments of student skills as they relate to word caller status. The proportion of teacher-nominated word callers and the obtained data were compared. Additionally, teachers' judgments about student performance and the usefulness of ORF were examined. Research questions were examined by grade level.

- Research Question 3a: Is there an association between teacher-nominated word callers and research-identified word callers in a sample of ELs?
- Research Question 3b: Is there an association between research-identified and teacher-nominated word callers when levels of English language proficiency are considered separately?
- Research Question 3c: Are there significant differences between teacher-nominated and research-identified word callers in terms of ORF and CST-ELA-RC performance?
- Research Question 4a: Are there significant differences among teacher-nominated word callers, poor readers, and comprehension proficient readers in terms of ORF and CST-ELA-RC performance?

- Research Question 4b: What is the proportion of teacher-nominated word callers, poor readers, and comprehension proficient readers for which teachers endorse the usefulness of ORF?

Methods

Participants

Student data were collected by school staff and provided to the author by school administrators. The obtained database contained demographic and testing information for a sample of third (N =199) and fifth (N = 196) grade ELs from three schools within a diverse, urban school district in Southern California. The majority of students were male for both third (n = 110) and fifth (n = 103) grade. Spanish was the primary language of all participants. Students attended a district where approximately 81.3% of students were receiving free or reduced lunch. Approximately 5.1% of the students in the sample were receiving special education services; however, all were missing data necessary for analysis (i.e., ORF and/or CST-ELA-RC scores) and could not be included.

Twenty-seven teachers from three schools were given information about the study and 24 (88%) chose to participate. The majority of teacher participants were female (n = 13) and all were responsible for instructing the majority of language arts activities for the student sample. Two teachers taught self-contained special education classes with third through fifth grade students and the remaining taught general education third (n = 12) or fifth (n = 10) grade classes. The sample comprised mainly Hispanic (30%), black (30%), and white teacher participants (21.7%). All teachers possessed certification in California to instruct ELs and several had obtained a master's level degree (n = 10). Teachers

ranged from 26 to 59 years in age (M = 40.9) and had been teaching from 3 to 13 years (M = 9.5).

Reading and Language Assessments

AIMSweb Reading Curriculum Based Measurement (R-CBM). R-CBM passages (Shinn & Shinn, 2002) are a measure of ORF. The passages were developed to correspond to grade level as calculated by the Fry (1968) readability formula. Thirty standardized passages are available for progress monitoring and three passages are available for screening at each grade level (second through eighth). Each passage is 300 words in length. Technical adequacy information was obtained from pilot student data on a larger pool of possible passages (Howe & Shinn, 2002). Passages were equated the following ways: alternate-form reliability, comparisons of means, standard deviations, and standard errors of measurement (SEM), and finally, readability formulas. A criterion for alternate-form reliability was set at .70; any passages that were below this criterion were excluded. Passages that exceeded plus or minus one SEM in terms of their mean were also excluded. Finally, passage difficulty was estimated with Lexile-graded standards (Stenner & Burdick; as cited by Howe & Shinn). Alternate-form reliability coefficients for the passages range from .81 to .90.

Administration of R-CBM is standardized: the examiner reads a set of directions before the student reads the passage aloud. Each passage is timed for one minute. Errors are mispronunciations, omissions, and hesitations or struggling for three seconds or longer. Errors are annotated on an examiner copy of the passages. The student's score is

recorded as the total of words read correctly (WRC); for the purposes of this study the median WRC score from three passages was used.

Evidence for the reliability and validity of ORF scores for ELs was described previously (e.g., Baker & Good, 1995; Muyskens et al., 2009; Wiley & Deno, 2005). However, no studies to date have examined the performance of ELs on AIMSweb passages specifically. For the purposes of this study ORF scores were converted into deviation standard scores (M = 100; SD =15) in order to calculate the word caller definition.

California Standards Tests -English Language Arts-Reading Comprehension (CST-ELA-RC). The CST -ELA-RC (ETS & CDE, 2009) is administered in the state of California to public school students in second through eleventh grade. The CST-ELA-RC is one subscale from the ELA domain from a group of tests (i.e., the CSTs) that also assesses math and science (in the elementary school grades). The test is given annually to document student achievement in subject areas that are aligned with state standards for grade level instruction. The CST-ELA-RC was chosen to represent reading comprehension achievement in the current study because of the importance placed on state-level tests for accountability (Good, Simmons, & Kame'enui, 2001; Reschly et al., 2009).

CST-ELA-RC selected items are released each year. The following information is based on the examination of the third and fifth grade released test items (CDE, 2009b). The items are based on paragraphs or short passages that are both fiction and non-fiction. Students are given several multiple-choice questions following the text that assess: recall

of details, making inferences, identifying the story sequence, making predictions, summarizing the story, and identifying the main idea.

ETS and the CDE (2009) reported that the 2008 administration of the CST-ELA-RC yielded internal consistency estimates (as measured by Cronbach's alpha) of .73 and .74 for third and fifth grade, respectively. Items on the CST were written by individuals with degrees in the content area being tested with a teaching and assessment background. Several internal reviews were conducted on test items with special attention to content validity, the difficulty of items using item response theory parameters, and sensitivity of items when ethnically and culturally diverse students are concerned (ETS & CDE, 2009). After initial item development and review, the validity of the CST-ELA-RC was further examined with expert content reviewers from ETS, CDE, and a panel of external educators that were not employed by either agency. To establish content validity, content area experts that held at minimum, a bachelor's degree in their field (most held advanced degrees) who also had extensive K-12 teaching and assessment experience examined test items for alignment with California state standards, clarity, grade-level and content area appropriateness, and format of presentation (ETS & CDE, 2009). Additionally, the entire CST-ELA was evaluated by comparing it to another, similar assessment to establish convergence. The California Achievement Test - Sixth Edition (CAT/6) Reading and Language tests were administered in conjunction with the CST-ELA; validity coefficients ranged from .75 to .80 (ETS & CDE, 2009). For the purposes of this study, CST-ELA-RC scores were converted into deviation standard scores ($M = 100$; $SD = 15$) in order to calculate the word caller definition.

California English Language Development Test (CELDT). The CELDT (CDE, 2009c) is a test used in the state of California to measure English language proficiency for purposes of educational planning. The test is given annually to all ELs and is purported to capture progress over time. The CELDT is comprised of four domains: Listening, Speaking, Reading, and Writing. Scores derived from the CELDT are posited to reflect English language proficiency and not academic achievement. The test is constructed in such a way that skills are not associated with age or grade level because students may be just beginning to learn English skills despite an advanced grade level or older age. The technical manual emphasizes that test content focuses on language development and not educational achievment (CDE, 2009c). The CELDT yields scores that reflect five different categories of English language proficiency: Beginning, Early Intermediate, Intermediate, Early Advanced, and Advanced. Scores are provided on this scale for each domain and for the overall score that reflects the combination of domains. A student is considered proficient in English when he or she attains at least Intermediate classification on all four domains; this corresponds to an overall CELDT score of Early Advanced or higher. Scores for third and fifth grade can also be reported in scale form from 230 to 700.

Reliability of the CELDT has been examined through measures of internal consistency, reported as coefficient alpha. Estimates across domains for third grade range from .73 - .86 and from .75 - .89 for fifth grade. Validity evidence for the CELDT has been demonstrated primarily through expert judgment, rather than through quantitative comparison to assessments that measure similar and different constructs. The CDE

(2009c) reported that the CELDT has been evaluated in terms of its alignment to the English Language Development Standards for California that it is designed to inform. Additionally, content appropriateness has been evaluated in terms of making determinations about whether items assess language ability versus academic achievement. Convergent validity evidence is demonstrated through intercorrelations amongst CELDT scales, these range from .49-.76 for third through fifth grade; test authors stated that there were no external measures to correlate with the CELDT (CDE, 2009c). Finally, the CELDT items were selected based on item response theory to evaluate item difficulty and discrimination. The overall CELDT score (which includes Listening, Speaking, Reading, and Writing domains) will be used for all analyses where CELDT is indicated, unless otherwise noted.

Procedure

Student data. The assessments used for the purposes of this study were part of the schools' routine assessments administered during the year. Student data were collected by school staff during the 2010-2011 year and included spring ORF, CELDT, and CST-ELA-RC. To include student data in the study, a passive consent and information letter was sent home to parents that described the purpose of the study, the data that would be used, and the option to decline. Three students' parents returned passive consent forms and declined inclusion of their child's data. These cases were removed from the database prior to analysis.

ORF data collection was conducted during spring of the academic year, by several of the classroom teachers (n =17, from two schools) and one reading specialist (who

collected data instead of the teachers at the other school) who were trained to administer the assessments in a standardized format. The author and research staff visited participating schools during the ORF screening period to measure administration fidelity using the *Accuracy of Implementation Rating Scale* (AIRS; Shinn & Shinn, 2002). The AIRS is a publisher-designed checklist to document standardization. Each staff member that was responsible for administering ORF was observed once at an unannounced visit for which they had given previous consent. Adherence to the AIRS was approximately 92.4% (N = 18). Inter-rater reliability on ORF scoring was also calculated during these visits. Reliability was calculated by dividing agreements by total words read in the passage and multiplying by 100. An average reliability of 98.38% (N = 18) was calculated.

Test administration occurred in a quiet place away from the students' classrooms and other visual and auditory distractions; this was usually in the hallway outside the classroom. Students received three grade level ORF passages individually and the median score was used.

CELDT and CST-ELA-RC data were collected by classroom teachers and specialists trained by the district to administer the tests in a standardized format. CELDT data were collected in the fall. CST-ELA-RC data were collected in the spring, primarily April and May.

Teacher Data. Teacher participants were given a packet containing two surveys to complete and return to the author. The *Teacher Survey* (see Appendix A) covered demographic information. The *Classroom Survey* (see Appendix B) documented teacher

judgment on the reading skills of students in his or her class. On the Classroom Survey, teachers were given a list of student participants from their class and asked to select one skill level for the categories "reading fluency" and "reading comprehension." Skill level for each category included four options: (a) "far below average", (b) "below average", (c) "average", and (d) above average." The ratings corresponded to descriptors of teacher-nominated profiles: *poor reader* ("below average" or "far below average" on fluency and comprehension)*, word caller* ("average" or "above average" on fluency and "below average" or "far below average" on comprehension)*,* and *comprehension proficient reader* (any rating on fluency and "average" or "above average" on comprehension). The comprehension proficient reader category thus includes two possible student profiles: students with adequate performance on fluency and comprehension in addition to students with below average teacher ratings on fluency and proficient or above ratings on comprehension. The profiles were grouped in this manner since differences in fluency performance for students that have proficient comprehension abilities are not the focus of the current study. An effort was made to avoid constraining teacher responses into any specific category in a way that might influence their choices. Therefore, teachers had the choice to select options that result in four different profiles even though data analysis only considered three (i.e., teacher-nominated poor readers, word callers, and comprehension proficient readers).

Next, teachers were asked to indicate whether ORF "is a useful assessment to screen this particular student for reading problems and determine if intervention is needed." The choice was dichotomous (i.e., "yes" or "no"). The definition of ORF,

adapted from two common publishers' descriptions (i.e., Good & Kaminski, 2002; Shinn & Shinn, 2002), was stated as:

> a task in which a student reads from a grade level appropriate passage of connected, meaningful text for a discrete time period. The examiner records the number of words read correctly in this time period. Omissions, mispronunciations, substitutions, and hesitations longer than three seconds are counted as incorrect. The total score is the number of words read correctly in one minute.

The author attended staff meetings at participating schools, explained the study, and passed out consent forms. Teachers that agreed to participate were contacted by a research assistant in the following weeks and given their surveys. Research assistants explained the directions on the surveys, asked teachers if they had any questions, and gave directions to return the surveys.

Classroom Survey Pilot Study. The Classroom Survey was piloted with a group of second through sixth grade teachers (N = 19) who rated ELs in their classrooms (N = 207). Teacher participants were primarily female (n = 11) and ranged from 36 to 58 years (M = 47.58). Teachers were White (75%) and Asian (16.6%) and on average had 16.9 years of teaching experience, ranging from 6 to 27 years. One teacher in the sample did not possess certification to instruct ELs in the state of California, but was still responsible for such instruction at her school. The purpose of the pilot was to determine if teachers would demonstrate variability in their responses when only four categories for rating fluency and reading comprehension were present on the survey. The pilot sought to determine if the majority of students would be rated "below average" or "far below

average," rendering poor variability. Results showed that teachers rated approximately 38% of the sample as "below average" or "far below average" in fluency and 45% "below average" or "far below average" for reading comprehension. This distribution of teacher responses was considered sufficient when four categories to describe student performance were used.

Word Caller Definitions

The current study defined at word callers defined two ways. ORF and CST-ELA-RC scores were both converted to the same standard score scale (M = 100, SD = 15) for purposes of comparison. For *Criterion 1*, word callers were defined by an ORF standard score greater than or equal to 95 and a standard score below 85 on the CST-ELA-RC. This definition replicates Meisinger and colleagues' (2009) cutoff scores. For *Criterion 2*, word callers were defined by a CST-ELA-RC standard score of at least two standard deviations or more below ORF. This definition is different since ORF scores did not have to be in the average to above average range (e.g., a student could have below average fluency and still be defined as a word caller if comprehension skills are sufficiently lower than ORF). This interpretation has yet to be examined and was included to determine if teachers' propensity to nominate word callers might be explained by large proportions of students with discrepancies between fluency and comprehension across all skill levels.

Results

Missing Data and Descriptive Analysis

Samples contained third (N= 199) and fifth (N = 196) grade students. However, several students' data were incomplete due to absence, moving away from the school,

teachers declining study participation, and teachers returning incomplete surveys. Missing data were handled with listwise deletion and were only deleted when the value(s) missing was required for a particular analysis. This procedure was selected based on recommendations from Allison (2000) and Scheffer (2002). Therefore, data was deleted in two stages.

First, students' cases were deleted if they were missing data required for research questions 1 and 2: CELDT, ORF, or CST-ELA-RC scores. The first round of deletion removed students from third (n = 18) and fifth (n = 21) grade samples. The remaining students were included in the analyses that addressed research questions 1 and 2. This left a sample of 181 third grade students (the majority were male, n = 100) and 175 fifth grade students (the majority were male, n = 89). Table 1 summarizes the sample mean, standard deviation, range, skewness, and kurtosis for third and fifth grade CELDT scale scores, ORF raw scores, ORF standard scores, CST-ELA-RC standard scores. Raw ORF means indicated performance at the 25th and 20th percentile for third and fifth grade, respectively. Average CELDT scores represent Intermediate levels of English language proficiency for both grades.

Second, the student cases that remained from the first round of deletion were deleted if they were missing any teacher judgment data that were required for research questions 3 and 4 (i.e., fluency rating, comprehension rating, ORF usefulness rating). The second round of deletion removed students from both third (n = 31) and fifth (n = 19) grade samples. The majority of the student cases were deleted due to two third grade teachers (n = 17) and one fifth grade teacher (n= 15) declining participation. The

remaining students were included in the analyses that addressed research questions 3 and 4. This left a sample of 150 third grade students (the majority were male, n = 84) and 156 fifth grade students (the majority were male, n = 80). Table 2 summarizes the sample mean, standard deviation, range, skewness, and kurtosis for third and fifth grade CELDT scale scores, ORF raw scores, ORF standard scores, CST-ELA-RC standard scores. Raw ORF means indicated performance at the 23rd and 22nd percentile for third and fifth grade, respectively. Average CELDT scores represent Intermediate levels of English language proficiency for both grades.

Descriptive statistics for teacher ratings of students' skills used for assignment to groups (word callers, poor readers, and comprehension proficient readers) are presented in Table 3. Teachers rated the majority of students as average in both reading fluency and comprehension for both grades, with the exception of fifth grade reading comprehension where the same proportion of students were rated as average and below average. ORF was rated as useful for the majority of students in both third and fifth grade.

Research Question 1

To address research question 1, multiple regression was used to test whether the relationship between ORF and CST-ELA-RC scores was moderated by CELDT level. The overall CELDT score representing all four domains was chosen for this analysis based on Abedi's (2008) recommendation to do so when domains are correlated; examination of the CELDT domains revealed this relationship across grades (CDE, 2009c).

A regression model was tested where ORF was the predictor, CELDT was the moderator, and CST-ELA-RC was the criterion variable. To address the possibility of moderation, an interaction term between ORF and CELDT was included in the model, as recommended by Agresti and Finlay (1997). The ORF and CELDT variables were centered (i.e., the mean was subtracted from every score) as recommended by Hoyt, Imel, and Chan (2008) to increase interpretability of regression coefficients. All terms were entered into the regression equation simultaneously since this approach yields identical results to a hierarchical model when only two predictors and a single interaction are included (Hoyt, Imel, & Chan, 2008). The following regression equation was used for third and fifth grade separately:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3\ x_1x_2 + e_i$$

where:

y = CST-ELA-RC scale score

b_0 = intercept

b_1 = standardized beta weight for ORF

b_2 = standardized beta weight for CELDT

b_3 = standardized beta weight for the interaction between ORF and CELDT

e_i = error

Assumptions of independence, linearity, normality, and homogeneity of variance were examined for both grades. All tests were independently measured and plots between each predictor and CST-ELA-RC revealed a linear relationship. Normality was examined through skewness and kurtosis; Marcoulides and Hershberger (1997) recommend values

between plus and minus one. Data summarized in Table 1 show that all variables, except one (third grade CELDT) satisfied this requirement. However, the distribution for third grade CELDT was only slightly leptokurtic (kurtosis = 1.19) and further examination of residual plots for homoskedasticity revealed that all assumptions were met. All bivariate correlations were significant and are reported in Table 4.

Results for the third grade interaction model indicated that the overall model was significant, $R^2 = .32$, $F\ (3,177) = 27.03$, $p < .01$. However, examination of the individual predictors revealed the interaction term between ORF and CELDT was not significant. In the case where the interaction term is not significant, Agresti and Finlay (1997) recommend it be removed from the regression equation and the model be tested to examine the partial effects of the other predictors. Thus, a non-interaction model was tested for third grade using only ORF and CELDT to predict CST-ELA-RC scores:

$$y = b_0 + b_1x_1 + b_2x_2 + e_i$$

where:

y = CST-ELA-RC scale score

b_0 = intercept

b_1 = standardized beta weight for ORF

b_2 = standardized beta weight for CELDT

e_i = error

The third grade non-interaction model was significant overall, $R^2 = .31$, $F\ (2,178) = 39.6$, $p < .01$. Change statistics revealed no significant differences between the interaction model and the non-interaction model for third grade, R^2 change = -.01, F

(1,177) = 1.72 p = .19. ORF and CELDT were significant terms in the model. Results of the regression analysis indicate the relationship between ORF and CST-ELA-RC for third grade does not change dependent on CELDT level. Summary of the regression analysis including beta weights and partial correlations for third grade is presented in Table 5.

A similar analysis was repeated for fifth grade. Results for the fifth grade interaction model indicated that the overall model was significant, R^2 = .34, F (3,171) = 28.68, p < .01. The individual terms for ORF and CELDT and the interaction term between ORF and CELDT were significant, thus the model was retained in its entirety. Regression results for fifth grade indicate that the relationship between ORF and CST-ELA-RC is moderated by CELDT. Figure 1 shows the regression lines for three categories of CELDT: Beginning/Early Intermediate, Intermediate, and Early Advanced/Advanced. The relationship between ORF and CST-ELA-RC depreciates as CELDT level decreases. Summary of the regression analysis including beta weights and partial correlations for fifth grade is presented in Table 6.

Research Question 2a

Research question 2a focused on the prevalence of word callers in the obtained sample. Proportions were calculated for third and fifth grade separately.

Criterion 1. For Criterion 1 a word caller was defined as any student whose ORF score was greater than or equal to a standard score of 95 for his or her grade level and whose corresponding CST-ELA-RC scale score was below 85. For third and fifth grade, respectively, 6% (n = 11) and 8% (n = 15) of the overall sample matched the Criterion 1 definition of word caller. There was no association between grade and Criterion 1 word

caller status, χ^2 (1, N = 361) = .687, p = .407, phi = .04. Coefficient phi represents effect size for chi-square with values of .10, .30, and .50 interpreted as small, medium, and large (Green & Salkind, 2005). A summary is presented at the bottom of Table 7. Results showed that some word callers do exist in EL samples using a definition similar to those used in past studies (i.e., Meisinger et al., 2009).

Descriptive statistics for CELDT, raw ORF scores, and CST-ELA-RC for Criterion 1 word callers are presented in Table 8. Third grade Criterion 1 word callers had an average Intermediate CELDT level. Their raw ORF scores ranged from the 38th percentile to the 69th percentile according to AIMSweb normative guidelines, and their average CST-ELA-RC scores were below average. Fifth grade Criterion 1 word callers had an average Intermediate CELDT level. Their raw ORF scores ranged from the 37th percentile to the 70th percentile according to AIMSweb normative guidelines, and their average CST-ELA-RC scores were below average.

Teacher judgment data indicated that most third grade Criterion 1 word callers were rated as average or above on both fluency (66.7%) and comprehension (55.5%). Most cases showed no discrepancy between fluency and comprehension ratings (n = 6, out of 9 that were rated). The nine Criterion 1 word callers with complete teacher judgment data represented the following teacher-nominated categories: five comprehension proficient readers, three poor readers and one word caller. Therefore, teachers had accurately rated the skills of one student from this group. ORF was rated as useful for 77.8% of the word callers.

For fifth grade, teacher judgment data showed that most Criterion 1 word callers were rated as average or above on fluency (60%) and close to half (46.7%) were rated average or above average on comprehension. Again, most cases showed no discrepancy between fluency and comprehension ratings (n = 9 out of 15 total). The fifteen Criterion 1 word callers represented the following teacher-nominated categories: seven comprehension proficient readers, six poor readers, and two word callers. Therefore, teachers had accurately rated the skills of two students in this group. ORF was rated as useful for 53.5% of the word callers.

Criterion 2. For Criterion 2, a word caller was defined as any student that has a CST-ELA-RC standard score equal to or greater than two standard deviations (30 standard score points) below his or her ORF standard score. Third and fifth grade were again examined separately. The prevalence of word callers that matched the Criterion 2 definition was 2% (n =3) for third and 1.1% (n = 2) for fifth grade. There was no association between grade and Criterion 2 word caller status, χ^2 (1, N = 360) = .141, p = .707, phi = .02. Results showed that few word callers exist among ELs when they are defined with high ORF and low CST-ELA-RC scores across all levels of performance. The information at the bottom of Table 7 provides a summary of these proportions.

Reliability of difference scores. The reliability of the difference scores used for both Criterion 1 and 2 word caller definitions were examined. Even when two tests have sufficient individual reliability, the reliability of their difference scores is affected negatively when the tests are correlated (Thorndike, 2005). It was determined from previous analyses (see Table 4) that ORF and CST-ELA-RC were correlated. The

reliability of the difference score for ORF and CST-ELA-RC was .56 for third and .48 for fifth grade. Given the low reliabilities, further examination was conducted that involved looking at each test's standard error of measurement (*SEM*) for each grade in order to determine which, if any, difference scores were not likely due to measurement error.

For both third and fifth grade the standard errors of measurement using a 68% confidence interval were equal to approximately eight for CST-ELA-RC scores and seven for ORF scores. This was not a problem for the Criterion 2 definition since difference scores needed to exceed two standard deviations, or 30 standard score points, in order to be considered a word caller. Even with scores on the lower end for ORF and higher end of the range for CST-ELA-RC, there would still be a 15-point difference. For Criterion 1, however, there was a 10-point difference between scores of 95 or above on ORF and below 85 on CST-ELA-RC. The 10-point difference could potentially be due to measurement error alone for some students. For example, if a student has an obtained ORF score of 96 and a CST-ELA-RC score of 84, the ranges associated with those scores are 89 to 103 for ORF and 76 to 92 for CST-ELA-RC; there is overlap between the ranges that would suggest that the student might not fit the word caller definition. Therefore, data were examined for third and fifth grade separately to determine the magnitude of the difference scores among the Criterion 1 word callers.

A summary of the magnitude of difference scores for third and fifth grade is presented in Table 9. For third grade, the average difference score for Criterion 1 word callers was greater than the minimum 10-point difference (M = 21.89). Further, a t-test using word caller status as the independent variable (non-word caller versus Criterion 1

word caller) showed that the difference scores of Criterion 1 word callers were significantly higher on average than non-word callers (M = -2.66), t (179) = -5.58, $p <$ .01.

For fifth grade, the average difference score for Criterion 1 word callers was also greater than the minimum 10-point difference (M = 22.85). Again, a t-test using word caller status as the independent variable (non-word caller versus Criterion 1 word caller) showed that the difference scores of Criterion 1 word callers were significantly higher on average than non-word callers (M = .80), t (173) = -7.25, $p < .01$.

It was concluded that despite low reliability of the difference scores, the students identified by both definitions of word callers had average difference scores that were larger than non-word callers and were likely due to factors beyond measurement error alone.

Research Question 2b

To address research question 2b, both word caller definitions were applied to the sample stratified by three levels of the CELDT: Beginning/Early Intermediate, Intermediate, and Early Advanced/Advanced) for both third and fifth grades. The results of this stratification are reported in Table 7.

For Criterion 1, Intermediate CELDT level students had greatest proportion of word callers for both grades (third = 11.3%; fifth = 9.5%). Third grade Early Advanced/Advanced CELDT level students had the smallest proportion (0%). For third grade, there was an association between CELDT level and Criterion 1 word caller status, χ^2 (2, N = 181) = 8.44, $p < .05$, phi = .22. Follow-up analysis using Holm's sequential

Bonferroni method (Green & Salkind, 2005) showed Intermediate CELDT level students had a greater proportion of word callers than Beginning/Early Intermediate students, χ^2 (1, N = 159) = 6.04, $p < .016$, phi = .20; other pairwise comparisons were not significant. In contrast, among fifth grade students it was the Beginning/Early Intermediate CELDT level that had the smallest proportion (7%) of word callers. However, there was no association between CELDT level and Criterion 1 word callers for fifth grade students, χ^2 (1, N = 156) = .158, $p = .924$, phi = .03.

For Criterion 2, proportions were very small across all categories for both grades and ranged from 0% to 2.4%. There was no association between Criterion 2 word caller status and CELDT level for third grade, χ^2 (2, N = 181) = .602, $p = .74$, phi = .06, or fifth grade, χ^2 (2, N = 156) = 1.97, $p = .37$, phi = .11.

Research Question 3a

Research question 3a addressed whether there was an association between teacher-nominated word callers and research-identified word callers (i.e., students who met the Criterion 1 word caller definition). Criterion 1 was chosen since it most closely matched the word caller definition for which teachers had nominated students in past studies (i.e., Hamilton & Shinn, 2003; Meisinger et al., 2009; Meisinger et al., 2010). Two levels were used for each teacher-nominated and research-identified source: word caller and non-word caller. Any student who was not a teacher-nominated word caller (i.e., poor readers and comprehension proficient readers) was placed in the non-word caller group.

A chi-square analysis was used for this question. Analyses for third and fifth grade were conducted separately, to examine whether there was an association between the research-identified and teacher-nominated word callers in the sample.

For the third grade sample, teacher-nominations of word callers were not associated with research-identifications, Pearson χ^2 (1, N = 150) = 0, p = .983, phi = .00. Similar results were found for the fifth grade sample: teacher-nominations of word callers were not associated with research-identifications, Pearson χ^2 (1, N = 156) = .171, p = .679, phi = .03. Teachers had nominated different students than those that were identified through research criteria.

Since chi-square analyses do not give information about the magnitude of accuracy, predictive accuracy indices were also calculated to determine how accurate teachers' nominations of word callers were in predicting research-identified word caller status. Predictive accuracy indices use four possible outcomes of any classification: valid positive, false positive, false negative, and valid negative (Rathvon, 2004). In this case, valid positive means that a teacher's ratings indicated the student was a word caller and the student subsequently emerged as a research-identified word caller, whereas false positive means that a teacher's ratings indicated the student was a word caller and the student did not emerge as a research-identified word caller. Valid negative means that a teacher's judgments indicated the student was not a word caller and the student did not emerge as a research-identified word caller, whereas false negative means that a teacher's judgments indicated the student was not a word caller and the student emerged as a research-identified word caller.

The predictive accuracy indices that evaluate the outcomes are named sensitivity, specificity, positive predictive value, negative predictive value, and hit rate. For this analysis, the degree to which teacher judgments accurately identify word callers is sensitivity. The degree to which teacher judgments accurately identify those who are not word callers is specificity. Positive predictive value refers to the proportion of students correctly nominated as word callers in comparison to the total number of students identified as word callers. Negative predictive value refers to the proportion of students correctly nominated as non-word callers in comparison to the total number of students identified as non-word callers. Finally, the hit rate refers to the proportion of students that were correctly nominated overall (i.e., valid positives and valid negatives). The metric by which predictive accuracy indices are judged is subjective as are the cutoff scores used to dichotomize performance on the predictor and outcome. Rathvon (2004) reported a consensus that sensitivity, specificity, and positive predictive value should all be equal to or greater than 75%.

For third grade, predictive accuracy indices showed that for the overall sample, teachers did a poor job of rating the skills of students that actually emerged as research-identified word callers; the false negative rate was high and resulted in a low sensitivity index of 11%. Additionally, teachers misjudged students as word callers; the false positive rate was high and resulted in a low positive predictive value of 6%. The overall hit rate for third grade was 84% because teachers were fairly accurate in rating the skills of non-word callers.

For fifth grade, predictive accuracy indices showed that for the overall sample, teachers again did a poor job of rating the skills of students that actually emerged as research-identified word callers; the false negative rate was high and thus had a low sensitivity index of 13%. Additionally, teachers misjudged students as word callers; the false positive rate was high and thus had a low positive predictive value of 13%. The hit rate for fifth grade was 83% because teachers were fairly accurate in rating the skills of non-word callers. The overall results for third and fifth grade are summarized in Tables 10 and 11, respectively. Results showed that for both third and fifth grade, teacher-nominations did not approach a high degree of accuracy for word callers.

Research Question 3b

To address research question 3b, the analysis for question 3a was repeated separately for third and fifth grade for each of three groups stratified by clusters of CELDT level: Beginning/Early Intermediate, Intermediate, and Early Advanced/Advanced. The results of the chi-square analyses and predictive accuracy indices for third and fifth grade stratified by CELDT level are again summarized in Tables 10 and 11, respectively.

Similar to the analysis of the entire sample of third and fifth grade, results of the chi-square analyses showed that the association between teacher-nominated and research-identified categories of word callers was not significant for any CELDT level cluster for either grade. Predictive accuracy indices were also poor and showed that sensitivity ranged from 0% to 50% and positive predictive value from 0% to 33%. Teachers' nominations of word callers did not match those identified by research criteria when

students were stratified by CELDT level clusters, nor did they approach an acceptable level of accuracy.

Research Question 3c

To address research question 3c, a multivariate analysis was conducted to determine if there were significant differences between teacher-nominated and research-identified word callers in terms of ORF and CST-ELA-RC performance. Hotelling's T^2 statistic (as described by Marcoulides & Hershberger, 1997) was used with source of identification as the independent variable (teacher-nominated versus research-identified) and ORF and CST-ELA-RC as the dependent variables. This analysis was used to determine if the teacher-nominated word callers performed differently than the research-identified word callers on ORF and CST-ELA-RC. The analyses were conducted separately for third and fifth grade. A multivariate analysis was conducted because of the previously reported significant correlations between ORF and CST-ELA-RC. The means and standard deviations for both third and fifth grade variables are reported in Table 12.

Multivariate assumptions were examined for third grade: normality, linearity, and homoskedasticity. All assumptions were met with the exception of the tests for homogeneity of variance and covariance matrices. Box's Test indicated significant differences in covariance matrices, F (3, 4347.62) = 3.76, p = .010. However, when sample sizes are unequal, this test is sensitive to minor deviations from normality. In light of evidence of all other assumptions being met, the analysis was considered robust.

Third grade results showed a significant effect, Hotelling's T^2 = .91, F (2, 21) = 9.52, $p < .01$, partial $\eta^2 = .48$. Follow-up examination of the main effects for each

dependent measure showed that research-identified word callers had significantly lower CST-ELA-RC scores than teacher-nominated word callers, F (1, 22) =14.53, $p < .01$, partial $\eta^2 = .40$, and that there was no significant difference between the groups on ORF. Results indicated that for third grade, teacher-nominated word callers were inaccurate because teachers had underestimated students' CST-ELA-RC scores for this group; when their actual scores were compared to the research-identified group, teacher-nominated word callers had higher average CST-ELA-RC scores than teachers had estimated.

For the fifth grade analysis, multivariate assumptions were examined first: normality, linearity, and homoskedasticity. Again, all assumptions were met with the exception the tests for homogeneity of variance and covariance matrices. Box's Test indicated significant differences in covariance matrices, F (3, 140813.21) = 3.51, p = .015. Again, when sample sizes are unequal, this test is sensitive to minor deviations from normality. In light of evidence of all other assumptions being met, the analysis was considered robust.

A significant effect was found overall for fifth grade, Hotelling's $T^2 = 1.38$, F (2, 24) = 16.57, $p < .01$, partial $\eta^2 = .58$. Follow-up examination of the main effects for each dependent measure showed that research-identified word callers had significantly lower CST-ELA-RC scores than teacher-nominated word callers, F (1, 25) = 4.46, $p < .05$, partial $\eta^2 = .15$. Additionally, research-identified word callers had significantly higher ORF scores than teacher-nominated word callers, F (1, 25) = 11.40, $p < .01$, partial $\eta^2 = .31$. Results indicated that for fifth grade, teacher-nominated word callers were

inaccurately nominated for two reasons. First, teachers had underestimated students' CST-ELA-RC scores and second, teachers had overestimated students' ORF scores.

Research Question 4a

Research question 4a addressed whether patterns of performance on ORF and CST-ELA-RC measures significantly differed among teacher-nominated: word callers, poor readers, and comprehension proficient readers. This analysis was similar to that conducted by Meisinger and colleagues (2009). A multivariate analysis of variance (MANOVA) was conducted with teacher-nominated classifications as the independent variable (i.e., word caller, poor reader, and comprehension proficient reader) and ORF and CST-ELA-RC as the dependent variables. The means and standard deviations for both third and fifth grade variables are reported in Table 13.

First, third grade data were examined. Multivariate assumptions of normality, linearity, and homoskedasticity were tested and no violations were found. Overall, there was a significant effect, Wilks' Lambda = .51, F (4, 292) = 28.87, $p < .01$, partial η^2 = .28. Follow-up, univariate ANOVAs showed main effects for both ORF, F (2, 147) = 62.88, $p < .01$, partial η^2 = .46, and CST-ELA-RC, F (2, 147) = 20.62, $p < .01$, partial η^2 = .22. Since main effects were significant, Scheffe's test was used to examine differences among poor readers, word callers, and comprehension proficient readers for both ORF and CST-ELA. Results showed that comprehension proficient readers demonstrated significantly higher ORF scores than both word callers and poor readers, $p < .05$, and that word callers demonstrated significantly higher ORF scores than poor readers, $p < .05$. When CST-ELA-RC was examined, poor readers had significantly lower CST-ELA-RC

scores than both word callers and comprehension proficient readers ($p < .05$). However, word callers and comprehension proficient readers did not have significantly different ORF scores, $p = .59$.

The ranking of teacher-nominated categories for third grade (from highest to lowest) using mean scores on ORF was: comprehension proficient readers, word callers, and poor readers. For CST-ELA-RC, comprehension proficient readers' and word callers' mean scores were similar and were higher than poor readers'.

Second, fifth grade data were examined. Multivariate assumptions of normality, linearity, and homoskedasticity were tested and no violations were found. Overall, there was a significant effect, Wilks' Lambda = .86, F (4, 304) = 6.10, $p < .01$, partial $\eta^2 = .07$. Follow-up, univariate ANOVAs showed main effects for both ORF, F (2, 153) = 12.24, $p < .01$, partial $\eta^2 = .14$, and CST-ELA-RC, F (2, 153) = 4.41, $p < .05$, partial $\eta^2 = .05$. Since main effects were significant, Scheffe's test was used to examine differences between each pair of poor readers, word callers, and comprehension proficient readers for both ORF and CST-ELA. Results showed that comprehension proficient readers demonstrated significantly higher ORF scores than poor readers, $p < .05$, and word callers demonstrated significantly higher ORF scores than poor readers, $p < .05$. When CST-ELA-RC was examined, poor readers had significantly lower CST-ELA-RC scores than comprehension proficient readers, $p < .05$. Other combinations of categories were not significant for ORF or CST-ELA-RC performance.

The ranking of teacher-nominated categories for fifth grade (from highest to lowest) using students' mean scores on ORF showed that comprehension proficient

readers and word callers has similar scores and poor readers' scores were significantly lower. For CST-ELA-RC, poor readers had a lower mean score than comprehension proficient readers; the mean score for word callers was not significantly different from any other category.

Research Question 4b

Research question 4b was addressed by calculating the proportion of teacher responses that indicated ORF is a useful measure for each teacher-nominated category: word callers, poor readers, and comprehension proficient readers. For third grade, teachers endorsed the usefulness of ORF for 80.56% of comprehension proficient readers (58 out of 72), 76.47% of word callers (13 out of 17), and 93.44% of poor readers (57 out of 61). There was no association between teacher-nominated categories and usefulness ratings, χ^2 (2) = 5.59, p = .06. For fifth grade, teachers endorsed the usefulness of ORF for 62.8% of comprehension proficient readers (44 out of 70), 43.7% of word callers (7 out of 16), and 86.9% of poor readers (60 out of 69). There was an association between teacher-nominated categories and usefulness ratings, χ^2 (2) = 16.74, p < .01, phi = .33. Follow-up analyses using the Holm's sequential Bonferroni method (Green & Salkind, 2005) showed that teachers rated ORF as useful for a greater proportion of poor readers than word callers, χ^2 (1) = 14.52, p < .016, phi = .41 and a greater proportion of poor readers than comprehension proficient readers, χ^2 (1) = 10.71, p < .025, phi = .28. Results showed with the exception of fifth grade teacher-nominated word callers, teachers endorsed the usefulness for the majority of students in all other categories.

Discussion

This study sought to examine word callers in a sample of third and fifth grade ELs. Specifically, research questions addressed two areas, (a) the prevalence of word callers and the consideration of English language proficiency as a moderating factor and (b) teachers' judgments of student reading skills, nomination of word callers and endorsement of ORF usefulness.

Prevalence of Word Callers

The first research question addressed whether there was an association between ORF and reading comprehension (as measured by CST-ELA-RC) and whether English language proficiency (as measured by CELDT) moderated such a relationship. Results indicated that for third grade ELs, ORF and CELDT significantly predicted reading comprehension. However, English language proficiency did not act as a moderator. Therefore, the relationship between ORF and reading comprehension was consistent across all levels of English language proficiency for the third grade sample.

In contrast, for fifth grade, CELDT moderated the relationship between ORF and reading comprehension. As English language proficiency decreased, the degree of the relationship between ORF and reading comprehension depreciated. Essentially, ORF did not predict reading comprehension outcomes as well when English language proficiency was at its lowest. One interpretation of this result is that there could be another moderator variable that explains the variance between ORF and reading comprehension that was not measured as part of this study for students that have lower levels of English proficiency. Vocabulary and oral language become more important in order to comprehend textual

information (Proctor et al., 2005; Nakamoto et al., 2007; Vellutino et al., 2007) as students get older and these changes might be pronounced since fifth grade text generally contains more complex vocabulary than third grade text. Although the CELDT includes items that relate to vocabulary and oral language, the content on the CELDT might not have been comprehensive enough to capture students' level of functioning.

However, an alternative explanation is that because the fifth grade CST-ELA-RC is intended to capture student reading comprehension at the fifth grade level, ELs with the lowest levels of English language proficiency might have shown limited variance, or scores that were concentrated at the low end of the distribution. As expected, examination of the data for this subsample revealed that their CST-ELA-RC scores were concentrated at the low end of the score range and had a smaller standard deviation than the other two CELDT level groups. It was suspected that the limited variance in the Beginning/Early Intermediate CELDT level group contributed to the breakdown in the relationship between ORF and CST-ELA-RC and should be interpreted cautiously. However, the emergence of CELDT as a possible moderator made it important to stratify subsequent analyses by proficiency level.

The prevalence of word callers was examined in the third and fifth grade EL student samples both as a whole grade level and stratified by English language proficiency level. Of the two definitions considered, Criterion 1 defined the largest proportions in both the third and fifth grade sample. About 6% of third and 8% of fifth grade EL students were defined as word callers according to Criterion 1. The Criterion 1 definition was similar to that used by other researchers (Meisinger et al., 2009) in that a

student with a standard score greater than or equal to 95 for ORF and a standard score less than 85 for reading comprehension was considered a word caller. The proportion of research-identified word callers did not differ significantly by grade. This is inconsistent with the work of Meisinger and colleagues that used NESs as participants and showed larger proportions of word callers in fifth grade. Approximately 1.82% of third and 9.78% of fifth grade NES students in the sample were defined as word callers (Meisinger et al.).

The Criterion 1 word callers that did emerge in both grades showed ORF performance that would not have been classified as "at-risk" according to AIMSweb normative guidelines since scores ranged from the 38th through the 69th percentile for third grade and the 37th through 70th percentile for fifth grade. Word callers in both grades demonstrated below average CST-ELA-RC scores and average CELDT scores in the Intermediate range. Teacher judgments for the majority Criterion 1 word callers showed no discrepancies between fluency and comprehension ratings. Further, 55.5% of third and 46.7% of fifth grade word callers were rated as average or above on comprehension by their teachers. Teachers endorsed the usefulness of ORF for the majority of this group. This information shows that for Criterion 1 word callers neither ORF screening data nor teacher judgments were sensitive to word callers' below average comprehension skills. Although it is for a relatively small subsample of children (no screening measure is completely accurate) it suggests the need to examine other available reading comprehension data when using ORF as a screening tool with English learners across grade levels.

In contrast, only 2% of third and 1.1% of fifth grade EL students were defined as word callers according to Criterion 2. The Criterion 2 definition was a reading comprehension score that was less than an ORF score by at least two standard deviations. The magnitude of the difference for this definition was set quite high (i.e., 2 standard deviations) and few word callers according to this definition emerged for either grade.

Finally, when word caller prevalence was examined within English proficiency levels, a larger proportion of Criterion 1 word callers emerged in the third grade Intermediate level. No association between CELDT level cluster and word caller status was found for fifth grade. Results showed that Criterion 1 word caller prevalence generally did not vary as a function of English language proficiency status. These results are somewhat inconsistent given the regression analysis that indicated English language proficiency did not moderate the relationship between ORF and reading comprehension for third grade and did moderate the same relationship for fifth grade.

Overall, results supported past work that has suggested ORF is related to reading comprehension outcomes for ELs (Muyskens et al., 2009; Wiley & Deno, 2005). However, this must be considered in the context that for some ELs (those that are older and have lower levels of English language proficiency) the relationship between ORF and reading comprehension might not be as strong. The breakdown in this relationship might be more pronounced because of vocabulary and oral language skills that are particularly low (i.e., Beginning/Early Intermediate CELDT level) and might suggest the decrease in the association between ORF and reading comprehension as students get older demonstrated in past studies (e.g., Jenkins & Jewell, 1993; Shinn et al., 1992; Yovanoff,

et al., 2005) occurs only for students that fit this profile since other literature has not replicated this breakdown (e.g., Reschly et al., 2009; Wood, 2006). More likely, the result might be due to the limited variance in CST-ELA-RC scores for the Beginning/Early Intermediate CELDT level group. Additionally, English language proficiency added significant contributions to both models, regardless of interaction effects. These results build a strong case for considering both English language proficiency and grade level when interpreting ORF scores for ELs.

This conclusion is further supported when the prevalence of word callers is considered. Profiles of the obtained Criterion 1 word callers suggest neither ORF nor teacher judgments were sensitive to their poor comprehension. Although the proportions of word callers were relatively small in the current sample, the proportions were inconsistent with the results from past studies with NESs that showed word callers existed in very small proportions in third grade and increased in fifth grade (Meisinger et al., 2009). The prevalence of word callers among ELs in this study generally supports the validity of using ORF scores as a screening measure for ELs, since some false negatives (word callers in this case) are expected with any screening measure (Rathvon, 2004). However, the importance of considering English language proficiency, grade level, and alternative sources of reading comprehension data when making instructional decisions using ORF scores for ELs is underscored.

Teacher Judgments

The second set of research questions addressed the accuracy of teacher judgments about reading skills (which were used to form word caller, comprehension proficient

reader, and poor reader categories) and second, teacher endorsement of ORF usefulness for specific students. Teacher related judgments are important because they are likely to influence teachers' selection of ORF for assessment and the use of information from ORF scores to enhance instructional decision-making (Begeny, Krouse, Brown, & Mann, 2011).

Past research has suggested that teachers nominate students who they believe read fluently but do not comprehend at a commensurate level (Hamilton & Shinn, 2003, Meisinger et al., 2009; Meisinger et al., 2010) as word callers. The proportion of teacher-nominated word callers, or students whose teacher ratings were below average for reading comprehension and average to above average on fluency, was compared to Criterion 1 research-identified word callers. Criterion 1 was chosen since it most closely matched the word caller definition for which teachers had nominated students in past studies (i.e., Hamilton & Shinn, 2003; Meisinger et al., 2009; Meisinger et al., 2010). Results of the analysis indicated that for both the third and fifth grade overall sample, there was no association between teacher-nominated and research-identified word callers. Subsequent analyses that sought to determine the magnitude of accuracy with which teachers had nominated word callers showed teachers' nominations did not approach the minimum level of accuracy recommended by Rathvon of .75 (2004). There was little overlap between students who were research-identified and teacher-nominated word callers.

In addition, the association between teacher-nominated and research-identified word callers was also compared across English language proficiency levels measured by

CELDT for both third and fifth grade. Results showed that teachers' word caller nominations were generally different than the Criterion 1 word callers across CELDT levels for both third and fifth grade. Sensitivity indices ranged from 0 to 50%. Teachers were the most accurate when considering fifth grade, Beginning/ Early Intermediate students. For this subset, positive predictive power was only 33% (which is still very low compared to the .75 criterion recommended by Rathvon, 2004), but was the highest relative to other groups.

Lack of association and low predictive accuracy indices suggest teachers were inaccurate in their judgments of students' skills that resulted in a word caller profile. Inaccuracy of teacher judgments about student skills is consistent across the word caller literature (e.g., Hamilton & Shinn; Meisinger et al., 2009; Meisinger et al., 2010). This is critical since inaccurate teacher judgments might influence instructional decision-making and could adversely affect student outcomes (Begeny et al., 2011).

Past research has shown that teachers are generally inaccurate when asked to make point level estimates of reading ability (Feinberg & Shapiro, 2003) and tended to overestimate performance for students that were low readers (Hamilton & Shinn, 2003). This is further supported by other studies that have shown teachers estimate the skills of the highest performing readers the most accurately (Begeny, Eckert, Montarello, & Storie, 2008; Begeny et al., 2011). Further analyses showed the areas in which teachers were inaccurate.

A comparison between teacher-nominated and research-identified word callers in terms of ORF and CST-ELA-RC performance helped to clarify the nature of teachers'

inaccurate judgments. Results showed that for third grade, teacher-nominated word callers did not differ from research-identified word callers on ORF. However, the comparison between average CST-ELA-RC scores showed that teacher-nominated word callers were significantly higher. For fifth grade, teacher-nominated word callers had both significantly higher ORF scores and significantly lower CST-ELA-RC scores as compared to those for research-identified word callers. The results show an inconsistent pattern in terms of overestimation and underestimation. There is however, alignment with past research suggesting teachers inaccurately judge the performance of average and low level readers (Begeny et al., 2008; Begeny et al., 2011).

Teacher judgments were explored further by comparing word callers, comprehension proficient readers and poor readers in terms of ORF and reading comprehension to determine what patterns emerged. For third grade, teacher-nominated word callers and comprehension proficient readers did not differ on CST-ELA-RC, whereas for fifth grade teacher-nominated word callers did not differ on ORF or CST-ELA-RC. This suggests across grades, teachers inaccurately judged the reading comprehension skills of the students that emerged as teacher-nominated word callers. Although there were inconsistent patterns in terms of statistical significance between groups, mean performance of each group on ORF and CST-ELA-RC followed a pattern in which poor readers were ranked lowest, word callers were in the mid-range, and comprehension proficient readers were ranked highest. This pattern was evident for both third and fifth grade. This is consistent with past literature that shows teachers are

accurate at rank ordering students and making judgments about skills relative to other students (Feinberg & Shapiro, 2003; Feinberg & Shapiro, 2008).

Finally, the last research question addressed the use of ORF scores and examined teacher endorsement of ORF usefulness for word callers, poor readers, and comprehension proficient readers. It was hypothesized that teachers might not endorse ORF usefulness for students who were teacher-nominated word callers. However, results showed that teachers were generally consistent across all three groups for both third and fifth grade, with the exception of fifth grade poor readers, who were given a larger proportion of ORF usefulness ratings compared to word callers and comprehension proficient readers. This is similar to the results from the Meisinger and colleagues study (2010) that showed teachers' recommendations of assessment methods were generally aligned with the skills they associated with word caller definitions. Teachers endorsed the usefulness of ORF for the majority of students in every category.

Overall, these results are encouraging because despite the prevalence of some word callers, ORF was related to CST-ELA-RC performance in this sample. Teachers' ratings might ultimately affect their use of ORF, their interpretation of ORF scores, and their application of ORF score information to academic planning and instructional decision making. It is concerning, however, that teachers were unable to discriminate for whom ORF would be most appropriate: teachers endorsed the usefulness of ORF for many of the students that emerged as Criterion 1 research-identified word callers. Specifically, teachers rated ORF as useful for 77.8% of third grade and 53.5% of fifth grade word callers.

Limitations

There were limitations to this study that must be noted. First, data were drawn from a sample of ELs with Spanish as a primary language. Therefore, results are limited in terms of generalization to other English learners. Second, data were drawn from a school district where the proportion of students using free and reduced price lunch was large, again limiting the generalization of results to students from low socio-economic backgrounds.

Next, the word caller definition is limited because it is arbitrary. Care was taken to identify meaningful cutoff scores based on past work (i.e., Meisinger et al., 2009), however, there might be other educationally relevant definitions using different cutoff scores.

There are also limitations in terms of using teacher participants to judge student skills. Teacher judgment ratings were drawn from 24 teachers that rated approximately 356 students total. Although each teacher rated multiple students, the analyses were conducted at the student level, making some sets of student ratings technically not independent of one another. Other studies have conducted analyses in a similar fashion (e.g., Begeny et al., 2011; Hamilton & Shinn, 2003; Meisinger et al., 2009; Meisinger et al., 2010) and have noted the same limitation (i.e., Begeny et al., 2011).

Finally, there was also a limitation in the way teacher judgments were recorded. Although the Teacher Survey was piloted with a small group of teachers prior to the study a true investigation of its psychometric properties was not possible. Additionally,

the categories derived from the survey were researcher-defined and arguably student performance could have been grouped in several different ways.

Implications and Future Directions

Future directions should focus on further examination of English language proficiency as a moderator of the relationship between ORF and reading comprehension using different measures than the CELDT to determine if results can be replicated. Additionally, an alternate English language proficiency measure would also be useful for replicating the results of word caller prevalence stratified by various English language proficiency levels since the current study demonstrated conflicting results. With this in mind, it is crucial that practitioners continue to consider English language proficiency and grade level when assessment is conducted since results indicated that different interpretations of ORF screening scores might be necessary for ELs. Practitioners should be aware that word callers do emerge, although in relatively infrequent circumstances and take care to examine all available sources of reading comprehension data when making decisions about risk status. Screening with ORF can be bolstered when its validity is suspect through the inclusion of an assessment like maze that provides an indication of reading comprehension that is valid for older students (Wayman et al., 2007).

Further study on the accuracy of teacher academic judgments is also necessary. Expanding on the work of Meisinger and colleagues (2010) and examining how teacher judgments apply to instructional decision-making would be useful. Teacher decision-making is used for crucial educational decisions (Begeny et al., 2011) and results of this study and others (Begeny et al., 2008; Begeny et al., 2011; Hamilton & Shinn, 2003;

Meisinger et al., 2009; Meisinger et al., 2010) have shown that teachers are inaccurate in this area. Results further bolster the need to share objective screening data with teachers so that subjective judgments are not necessary. Ultimately, providing professional development on the general inaccuracy of skill judgment and ways to circumvent this problem through data might be a decent endeavor for practitioners and an interesting subject for research to determine its effectiveness.

References

Abedi, J. (2004). The No Child Left Behind Act and English Language Learners: Assessment and Accountability Issues. *Educational Researcher, 33* (1), 4-14.

Abedi, J. (2007). English Language Proficiency Asssessment and Accountability under NCLB Title III: An Overview. In J. Abedi, *English Language Proficiency Assessment in the Nation: Current Status and Future Practice* (pp. 3-12). Davis, CA: The Regents of the University of California.

Abedi, J. (2008). Measuring Students' Level of English Proficiency: Educational Significance and Assessment Requirements. *Educational Assessment*, 13, 193-214.

Allison, P.D. (2000). Multiple Imputation for Missing Data: A Cautionary Tale. *Sociological Methods and Research, 28* (3), 301-309.

Al Otaiba, S., Petscher, Y., Pappamihiel, N., Williams, R., Dyrlund, A., & Connor, C. (2009). Modeling Oral Reading Fluency Development in Latino Students: A Longitudinal Study Across Second and Third Grade. *Journal of Educational Psychology, 101* (2), 315-329.

Albers, C., Kenyon, D., & Boals, T. (2009). Measures for Determining English Language Proficiency and the Resulting Implications for Instructional Provision and Intervention. *Assessment for Effective Intervention, 34* (2), 74-85.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA/APA/ACME). (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The Critical Role of Vocabulary Development for English Language Learners. *Learning Disabilities Research & Practice, 20* (1), 50-57.

Baker, S., & Good, R. (1995). Curriculum-Based Measurement of English Reading with Bilingual Hispanic Students: A Validation Study with Second-Grade Students. *School Psychology Review, 24* (4), 561-578.

Baker, S., Smolkowski, K., Katz, R., Fien, H., Seeley, J., Kame'enui, E., et al. (2008). Reading Fluency as a Predictor of Reading Proficiency in Low-Performing, High-Poverty Schools. *School Psychology Review, 37* (1), 18-37.

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.

Bradley, A.P. & Longstaff, I.D. (2004). Sample size estimation using the receiver operating characteristic curve. *Proceedings on the 17th International Conference on Pattern Recognition* 4, 428-431.

Burd, L., Kerbeshian, J., & Fisher, W. (1985). Inquiry into the incidence of hyperlexia in a state- wide population of children with pervasive developmental disorder. *Psychological Reports*, 57, 236-238.

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.

California Department of Education. (2009a). *DataQuest.* Retrieved March 4, 2010, from http://data1.cde.ca.gov/dataquest/

California Department of Education. (2009b). English Language Arts Released Test Questions. Retrieved May 5, 2011, from http://www.cde.ca.gov/ta/tg/sr/css05rtq.asp.

California Department of Education. (2009c). *Technical Report for the California English Language Development Test (CELDT).* Monterey: California Department of Education by CTB/McGraw-Hill LLC.

Capps, R., Fix, M., Murray, J., Ost, J., Passel, J., & Herwantoro, S. (2005). *The New Demography of America's Schools: Immigration and the No Child Left Behind Act.*Washington, DC: The Urban Institute.

Carlo, M., August, D., McLaughlin, B., Snow, C., Dressler, C., Lippman, D., et al. (2004). Closing the gap: Addressing the vocabulary needs of English-language leaners in bilingual and mainstream classrooms. *Reading Research Quarterly, 39* (2), 188-215.

Catts, H.W. & Hogan, T.P. (2003). Language Basis of Reading Disabilities and Implications for Early Identification and Remediation. *Reading Psychology*, 24, 223-246.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, New Jersey: Lawrence Erlbaum Publishers.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin,* 112(1), 155-159.

Children's Educational Services. (1987). *Standard reading passages*. Eden Prairie, MN: Author.

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.

D'Angiulli, A., Siegel, L., & Maggi, S. (2004). Literacy Instruction, SES, and Word-Reading Achievement in English-Language Learners and Children with English as a First Language: A Longitudinal Study. *Learning Disabilities Research & Practice, 19* (4), 202-213.

Deno, S. (2003). Developments in Curriculum-Based Measurement. *The Journal of Special Education, 37* (3), 184-192.

Deno, S., Maruyama, G., Espin, C., & Cohen, C. (1990). Educating students with mild disabilities in general education classrooms: Minnesota alternatives. *Exceptional Children*, 49, 36-45.

Dewitz, P. & Dewitz, P.K. (2003). They Can Read the Words, but They Can't Understand: Refining Comprehension Assessment. *The Reading Teacher*, 56(5), 422-435.

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

Educational Testing Service, California Department of Education, Standards and Assessment Division. (2009, March). *California Standards Tests: Technical Report: Spring 2008 Administration.* Retrieved March 4, 2010 from: http://www.cde.ca.gov/ta/tg/sr/documents/csttechrpt08.pdf

Feinberg, A., & Shapiro, E. (2003). Accuracy of Teacher Judgments in Predicting Oral Reading Fluency. *School Psychology Quarterly, 18* (1), 52-65.

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal, 21*(2), 449-460.

Foegen, A., Espin, C., Allinder, R., & Markell, M. (2001). Translating Research Into Practice: Preservice Teachers' Beliefs About Curriculum-Based Measurement. *The Journal of Special Education, 34* (4), 226-236.

Francis, D., & Rivera, M. (2007). Principles Underlying English Language Proficiency Tests and Academic Accountability for ELLs. In J. Abedi, *English Language Proficiency Assessment in the Nation: Current Status and Future Practice* (pp. 13-32). Davis, CA: The Regents of the University of California.

Fry, E. (1968). A readability formula that saves time. *Journal of Reading*, 11, 513-516.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*(2), 188-192.

Fuchs, L.S., & Deno, S.L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57*(6), 488-501.

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal, 21*(2), 449-460.

Fuchs, L., & Fuchs, D. (1986). Effects of Systematic Formative Evaluation: A Meta-Analysis. *Exceptional Children, 53* (3), 199-208.

Fuchs, L., Fuchs, D., & Hamlett, C. (1989a). Effects of Alternative Goal Structures Within Curriculum-Based Measurement. *Exceptional Children, 55* (5), 429-438.

Fuchs, L., Fuchs, D., & Hamlett, C. (1989b). Effects of Instrumental Use of Curriculum-Based Measurement to Enhance Instructional Programs. *Remedial and Special Education, 10* (2), 43-52.

Fuchs, L., Fuchs, D., Hosp, M., & Jenkins, J. (2001). Oral Reading Fluency as an Indicator of Reading Competence: A Theoretical, Empirical, and Historical Analysis. *Scientific Studies of Reading, 5* (3), 239-256.

Gersten, R., Baker, S., Shanahan, T., Linan-Thompson, S., Collins, P., & Scarcella, R. (2007). *Effective Literacy and English Language Instruction for English Learners in the Elementary Grades: A Practice Guide (NCEE 2007-4011).* Washington, DC: National Center for Education Evaluation and Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides/.

Gersten, R., Compton, D., Connor, C.M., Dimino, J., Santoro, L., Linan-Thompson, S., and 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 from http://ies.ed.gov/ncee/wwc/publications/practiceguides/.

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

Goffreda, C., Diperna, J., & Pedersen, J. (2009). Preventive Screening for Early Readers: Predictive Validity of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). *Psychology in the Schools, 46* (6), 539-552.

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

Good, R.H., Kaminski, R.A., Smith, S., Laimon, D., & Dill, S. (2001). *Dynamic Indicators of Basic Early Literacy Skills* (5th ed.). Eugene: University of Oregon.

Good, R., Simmons, D., & Kame'enui, E. (2001). The Importance and Decision-Making Utility of a Continuum of Fluency-Based Indicators of Foundational Reading Skills for Third-Grade High-Stakes Outcomes. *Scientific Studies of Reading , 5* (3), 257-288.

Gough, P., Hoover, W., & Peterson, C. (1996). Some Observations on a Simple View of Reading. In C. Cornoldi, & J. Oakhill, *Reading Comprehension Difficulties* (pp. 1-13). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Gottardo, A., Collins, P., Baciu, I., & Gebotys, R. (2008). Predictors of Grade 2 Word Reading and Vocabulary Learning from Grade 1 Variables in Spanish-Speaking Children: Similarities and Differences. *Learning Disabilities Research and Practice*, 23(1), 11-24.

Gottardo, A. & Mueller, J. (2009). Are First- and Second- Language Factors Related in Predicting Second-Language Reading Comprehension? A Study of Spanish-Speaking Children Acquiring English as a Second Language from First to Second Grade. *Journal of Educational Psychology*, 101(2), 330-344.

Graves, A., Plasencia-Peinado, J., Deno, S., & Johnson, J. (2005). Formatively Evaluating the Reading Progress of First-Grade English Learners in Multiple-Language Classrooms. *Remedial and Special Education, 26* (4), 215-225.

Green, S.B. & Salkind, N.J. *Using SPSS for Windows and Macintosh: analyzing and understanding data.* Upper Saddle River, New Jersey: Pearson.

Grigorenko, E.L., Klin, A., Pauls, D.L., Senft, R., Hooper, C., & Volkmar, F. (2002). A Descriptive Study of Hyperlexia in a Clinically Referred Sample of Children With Developmental Delays. *Journal of Autism and Developmental Disorders*, 32(1), 3-12.

Hamilton, C., & Shinn, M. (2003). Characteristics of Word Callers: An Investigation of the Accuracy of Teachers' Judgments of Reading Comprehension and Oral Reading Skills. *School Psychology Review, 32* (2), 228-240.

Harcourt Brace & Co. (1997). *Stanford Achievement Test Series- Ninth Edition: Technical data report*. San Antonio, TX: Author.

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

Hintze, J., & Silberglitt, B. (2005). A Longitudinal Examination of the Diagnostic Accuracy and Predictive Validity of R-CBM and High-Stakes Testing. *34* (3), 372-386.

Hintze, J., Callahan, J., Matthews, W., Williams, S., & Tobin, K. (2002). Oral Reading Fluency and Prediction of Reading Comprehension in African American and Caucasian Elementary School Children. *School Psychology Review, 31* (4), 540-553.

Hoover, W., & Gough, P. (1990). The Simple View of Reading. *Reading and Writing: An Interdisciplinary Journal, 2*, 127-160.

Hosp, M., & Fuchs, L. (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., & Shinn, M. (2002). *Standard Reading Assessment Passages (RAPs) For Use in General Outcome Measurement: A Manual Describing Development and Technical Features.* Ede Prairie: Edformation.

Hoyt, W.T., Imel, Z.E., & Chan, F. (2008). Multiple Regression and Correlation Techniques: Recent Controversies and Best Practices. *Rehabilitation Psychology, 53* (3), 321-339.

Jenkins, J., Fuchs, L., van den Broek, P., Espin, C., & Deno, S. (2003). Sources of Individual Differences in Reading Comprehension and Reading Fluency. *Journal of Educational Psychology, 95* (4), 719-729.

Jenkins, J.R., Heliotis, J., Haynes, M., Stein, M., & Beck, K. (1986). Does "passive learning" account for disabled readers' comprehension deficits in ordinary reading situations? *Learning Disability Quarterly*, 9(1), 69-76.

Jenkins, J., Hudson, R., & Johnson, E. (2007). Screening for At-Risk Readers in a Response to Intervention Framework. *School Psychology Review, 36* (4), 582-600.

Jenkins, J., & Jewell, M. (1993). Examining the Validity of Two Measures for Formulative Teaching: Reading Aloud and Maze. *Exceptional Children, 59* (5), 421-432.

Johnson, E., Jenkins, J., Petscher, Y., & Catts, H. (2009). How Can We Improve the Accuracy of Screening Instruments? *Learning Disabilities Research & Practice, 24* (4), 174-185.

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

Kamii, C., & Manning, M. (2005). Dynamic Indicators of Basic Early Literacy Skills (DIBELS): A Tool for Evaluating Student Learning? *Journal of Research in Childhood Education, 20* (2), 75-90.

Karlsen, B. & Gardner, E. (1985). *Stanford diagnostic reading test* (3rd ed.). San Antonio, TX: Psychological Corp.

Kieffer, M. (2008). Catching Up or Falling Behind? Initial English Proficiency, Concentrated Poverty, and the Reading Growth of Language Minority Learners in the United States. *Journal of Educational Psychology, 100* (4), 851-868.

Kindler, A. (2002). *Survey of the States' Limited English Proficient Students and Available Educational Programs and Services 2000-2001 Summary Report.* Washington, DC: National Clearinghouse for English Language Acquisition & Language Instruction Educational Programs.

Klein, J., & Jimerson, S. (2005). Examining Ethnic, Gender, Language, and Socioeconomic Bias in Oral Reading Fluency Scores among Caucasian and Hispanic Students. *School Psychology Quarterly, 20* (1), 23-50.

Kranzler, J., Miller, M., & Jordan, L. (1999). An Examination of Racial/Ethnic and Gender Bias on Curriculum-Based Measurement in Reading. *School Psychology Quarterly, 14* (3), 327-342.

LaBerge, D., & Samuels, S. (1974). Toward a Theory of Automatic Information Processing in Reading. *Cognitive Psychology, 6*, 293-323.

Lesaux, N., & Siegel, L. (2003). The development of reading in children who speak English as a second language. *Developmental Psychology, 39*, 1005-1019.

Linan-Thompson, S., Cirino, P., & Vaughn, S. (2007). Determining English Language Learners' Response to Intervention: Questions and Some Answers. *Learning Disability Quarterly*, *30*, 185-195.

Linan-Thompson, S., Vaughn, S., Prater, K., & Cirino, P. (2006). The Response to Intervention of English Language Learners at Risk for Reading Problems. *Journal of Learning Disabilities, 39* (5), 390-398.

Lomax, R.G. (2001). *Statistical Concepts: A Second Course for Education and Behavioral Sciences* (2nd ed.) Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Lovett, M., De Palma, M., Frijters, J., Steinbach, K., Temple, M., Benson, N., et al. (2008). Interventions for Reading Difficulties: A Comparison of Response to Intervention by ELL and EFL Struggling Readers. *Journal of Learning Disabilities, 41* (4), 333-352.

MacGintie, W.H., MacGintie, R.K., Maria, K., Dreyer, L.G., & Hughes, K.E. (2000). *Gates-MacGintie reading tests* (4th ed.). Rolling Meadows, IL: Riverside.

Marcotte, A., & Hintze, J. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology, 47*, 315-335.

Marcoulides, G.A. & Hershberger, S.L. (1997). *Multivariate Statistical Methods: A First Course*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

McGlinchey, M., & Hixson, M. (2004). Using Curriculum-Based Measurement to Predict Performance on State Assessments in Reading. *School Psychology Review, 33* (2), 193-203.

McLoyd, V. (1998). Socioeconomic Disadvantage and Child Development. *American Psychologist, 53* (2), 185-204.

Meisinger, E., Bradley, B., Schwanenflugel, P., & Kuhn, M. (2010). Teachers' Perceptions of Word Callers and Related Literacy Concepts. *School Psychology Review, 39* (1), 54-68.

Meisinger, E., Bradley, B., Schwanenflugel, P., Kuhn, M., & Morris, R. (2009). Myth and Reality of the Word Caller: The Relation Between Teacher Nominations and Prevalence Among Elementary School Children. *School Psychology Quarterly, 24* (3), 147-159.

Messick, S. (1994). Foundations of Validity: Meaning and Consequences in Pyschological Assessment. *European Journal of Psychological Assessment, 10* (1), 1-9.

Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice, Winter*, 5-8.

Muller, K.E.; LaVange, L.M.; Ramey, S.L.; & Ramey, C.T. (1992). Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications. *Journal of the American Statistical Association*, *87*(420), 1209-1226.

Muyskens, P., Betts, J., Lau, M., & Marston, D. (2009). Predictive Validity of Curriculum-Based Measures in the Reading Assessment of Students who are English Language Learners. *The California School Psychologist, 14*, 11-21.

Nakamoto, J., Lindsey, K., & Manis, F. (2007). A longitudinal analysis of English language learners' word decoding and reading comprehension. *Reading and Writing, 20*, 691-719.

Nathan, R., & Stanovich, K. (1991). The Causes and Consequences of Differences in Reading Fluency. *Theory into Practice, 30* (3), 176-184.

National Clearinghouse for English Language Acquisition. (2007). *The Growing Numbers of Limited English Proficient Students.* Washington, DC: United States Deparment of Education: Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students.

National Reading Panel. (2000). *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction.* Washington, DC: National Institute of Childe Health and Human Development.

Newman, T.M., Macomber, D., Naples, A.J., Babitz, T., Volkmar, F., & Grigorenko, E.L. Hyperlexia in Children with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, 37, 760-774.

Parker, R., Hasbrouck, J.E., & Tindal, G. (1992). The Maze as a Classroom-Based Reading Measure: Construction Methods, Reliability, and Validity. *The Journal of Special Education, 26* (2), 195-218.

Pearce, L., & Gayle, R. (2009). Oral Reading Fluency as a Predictor of Reading Comprehension With American Indian and White Elementary Students. *School Psychology Review, 38* (3), 419-427.

Posner, M.I. & Snyder, C.R.R. (2004). Attention and Cognitive Control. In D.A. Balota & E.J. Marsh (Eds.), *Cognitive Psychology: Key Readings in Cognition*. (pp. 205-223). New York: Psychology Press.

Pray, L. (2005). How Well Do Commonly Used Language Instruments Measure English Oral-Language Proficiency. *Bilingual Research Journal, 29* (2), 387-409.

Proctor, C., Carlo, M., August, D., & Snow, C. (2005). Native Spanish-Speaking Children Reading in English: Toward a Model of Comprehension. *Journal of Educational Psychology, 97* (2), 246-256.

Rathvon, N. (2004). *Early Reading Assessment.* New York: The Guilford Press.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical Linear Models.* Thousand Oaks, CA: Sage Publications, Inc.

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

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

Roehrig, A., Petscher, Y., Nettles, S., Hudson, R., & Torgesen, J. (2008). Accuracy of the DIBELS Oral Reading Fluency Measure for Predicting Third Grade Reading Comprehension Outcomes. *Journal of School Psychology, 46*, 343-366.

Salvia, J., Ysseldyke, J., & Bolt, S. (2007). *Assessment in Special and Inclusive Education.* Boston: Houghton Mifflin Company.

Samuels, S. (2007). The DIBELS Tests: Is Speed of Barking at Print What we Mean by Reading Fluency? *Reading Research Quarterly, 42* (4), 563-565.

Scheffer, J. (2002). Dealing with Missing Data. *Research Letters in the Information and Mathematical Sciences, 3*, 153-160.

Shankweiler, D., Lundquist, E., Katz, L., Stuebig, K.K., Fletcher, J.M., Brady, S., et al. (1999). Comprehension and Decoding: Patterns of Association in Children with Reading Difficulties. *Scientific Studies of Reading*, 3(1), 69-94.

Shapiro, E., Keller, M., Lutz, J., Santoro, L., & Hintze, J. (2006). Curriculum-Based Measures and Performance on State Assessment and Standardized Tests. *Journal of Psychoeducational Assessment, 24* (1), 19-35.

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* (4th ed.). Boca Raton, FL: Chapman & Hall.

Shinn, M., & Bamonto, S. (1998). Advanced Applications of Curriculum-Based Measurement: "Big Ideas" and Avoiding Confusion. In M. Shinn, *Advanced Applications of Curriculum-Based Measurement* (pp. 1-31). New York: The Guilford Press.

Shinn, M. R., Good, R. H., Knutson, N., & Tilly, W. D. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *SchoolPsychology Review*, 21(3), 459-479.

Shinn, M., & Shinn, M. (2002). *AIMSweb Training Workbook.* Eden Prairie: Edformation.

Silberglitt, B., Burns, M., Madyn, N., & Lail, K. (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.

Snow, C., & Kim, Y. (2007). Large Problem Spaces: The Challenge of Vocabulary for English Language Learners. In R. Wagner, A. Muse, & K. Tannenbaum, *Vocabulary Acquisition: Implications for Reading Comprehension* (pp. 123-131). New York: Guilford Press.

Solari, E., & Gerber, M. (2008). Early Comprehension Instruction for Spanish-Speaking English Language Learners: Teaching Text-Level Reading Skills While Maintaining Effects on Word-Level Skills. *Learning Disabilities Research and Practice, 23* (4), 155-168.

Stage, S., & Jacobsen, M. (2001). Predicting Student Success on a State-mandated Performance-based Assessment Using Oral Reading Fluency. *School Psychology Review, 30* (3), 407-419.

Stothard, S.E. & Hulme, C. (1992). Reading comprehension difficulties in children: The role of language comprehension and working memory skills. *Reading and Writing: An Interdisciplinary Journal*, 4, 245-256.

Stanovich, K. (1986). Matthe Effects in Reading: Some Consequences of Individual Differences in the Acquisition of Literacy. *Reading Research Quarterly, 21* (4), 360-407.

Stecker, P., Fuchs, L., & Fuchs, D. (2005). Using Curriculum-Based Measurement to Improve Student Achievement: Review of Research. *Psychology in the School, 42* (8), 795-819.

Torgesen, J. (2002). The Prevention of Reading Difficulties. *Journal of School Psychology , 40* (1), 7-26.

Torgesen, J., Alexander, A., Wagner, R., Rashotte, C., Voeller, 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), 33-58.

U.S. Department of Education, National Center for Education Statistics, Institute of Education Sciences. (2004). *English Language Learner Students in U.S. Public Schools: 1994 and 2000.* U.S. Department of Education.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) Reading Assessments. (2007). *The Nation's Report Card.* U.S. Department of Education.

Vanderwood, M., Linklater, D., & Healy, K. (2008). Predictive Accuracy of Nonsense Word Fluency for English Language Learners. *School Psychology Review, 37* (1), 5-17.

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.

Vaughn, S., Linan-Thompson, S., Mathes, P., Cirino, P., Carlson, C., Pollard-Durodola, S., et al. (2006). Effectiveness of an English intervention for first-grade English language learners at risk for reading problems. *The Elementary School Journal, 107*, 153-180.

Vellutino, F., Tunmer, W., Jaccard, J., & Chen, R. (2007). Components of Reading Ability: Multivariate Evidence for a Convergent Skills Model of Reading Development. *Scientific Studies of Reading, 11* (1), 3-32.

Wanzek, J., & Vaughn, S. (2007). Research-Based Implications From Extensive Early Reading Interventions. *School Psychology Review, 36* (4), 541-561.

Wayman, M., Wallace, T., Wiley, H., Ticha, R., & Espin, C. (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.

Wiley, H., & Deno, S. (2005). Oral Reading and Maze Measures as Predictors of Success for English Learners on a State Standards Assessment. *Remedial and Special Education, 26* (4), 207-214.

Wood, D. (2006). Modeling the Relationship Between Oral Reading Fluency and Performance on a Statewide Reading Test. *Educational Assessment, 11* (2), 85-104.

Woodcock, R. (1987). *The Woodcock reading mastery tests.* Circle Pines, MN: American Guidance Service.

Yovanoff, P., Duesbery, L., Alonzo, J., & Tindal, G. (2005). Grade-Level Invariance of a Theoretical Causal Structure Predicting Reading Comprehension With Vocabulary and Oral Reading Fluency. *Educational Measurement: Issues and Practice , Fall*, 4-12.

Table 1

Descriptive Statistics for Research Questions 1 and 2

Variable	M	SD	Skewness	Kurtosis	Range
CELDT					
Third	466.08	47.83	-0.44 (.18)	1.19 (.35)	272-580
Fifth	521.49	38.48	-0.25 (.18)	0.22 (.37)	424-619
ORF Raw					
Third	91.96	42.12	-0.13 (.18)	-0.65 (.36)	4-199
Fifth	109.41	37.38	-0.12 (.18)	-0.11 (.37)	17-216
ORF Standard					
Third	90.21	14.69	-0.13 (.18)	-0.65 (.36)	59.53-127.56
Fifth	87.18	12.74	-0.12 (.18)	-0.11 (.36)	55.68-123.52
CST-ELA-RC					
Third	91.38	15.10	-0.00 (.18)	-0.94 (.36)	59.69-121.95
Fifth	84.49	13.46	0.27 (.18)	-0.51 (.37)	59.35-120.82

Note. Third grade, N = 181. Fifth grade, N = 175. Skewness and kurtosis values are followed by their standard error in parentheses. CELDT = California English Language Development Test. ORF Raw = Raw oral reading fluency scores as measured by AIMSweb Reading-Curriculum Based Measurement. ORF Standard = Standardized oral reading fluency scores as measured by AIMSweb Reading-Curriculum Based Measurement. CST-ELA-RC = Standardized California Standards Test- English Language Arts- Reading comprehension scores.

Table 2

Descriptive Statistics for Research Questions 3 and 4

Variable	M	SD	Skewness	Kurtosis	Range
CELDT					
Third	465.90	50.76	-0.41 (.20)	0.91 (.40)	272-580
Fifth	524.15	36.29	-0.28 (.20)	0.04 (.39)	430-619
ORF Raw					
Third	90.09	43.41	-0.11 (.20)	-0.74 (.40)	4-199
Fifth	112.67	35.89	-0.09 (.20)	0.04 (.39)	17-216
ORF Standard					
Third	89.56	15.14	-0.11 (.20)	-0.74 (.40)	59.53-127.56
Fifth	88.29	12.24	-0.09 (.20)	0.04 (.40)	55.68-123.52
CST-ELA-RC					
Third	91.19	15.14	-0.11 (.20)	-0.93 (.40)	59.69-121.95
Fifth	85.45	13.11	0.17 (.20)	-0.68 (.40)	59.35-116.90

Note. Third grade, N = 150. Fifth grade, N = 156. Skewness and kurtosis values are followed by their standard error in parentheses. CELDT = California English Language Development Test. ORF Raw = Raw oral reading fluency scores as measured by AIMSweb Reading-Curriculum Based Measurement. ORF Standard = Standardized oral reading fluency scores as measured by AIMSweb Reading-Curriculum Based Measurement. CST-ELA-RC = Standardized California Standards Test- English Language Arts- Reading comprehension scores.

Table 3

Descriptive Statistics for the Classroom Survey

	Third Grade		Fifth Grade	
Variable	Frequency	Percent	Frequency	Percent
Reading Fluency Rating				
Above Average	23	15.3	9	5.8
Average	60	40.0	64	41.0
Below Average	29	19.3	56	35.9
Far Below Average	38	25.5	27	17.0
Reading Comprehension Rating				
Above Average	18	12.0	4	2.6
Average	4	36.0	64	41.0
Below Average	42	28.0	64	41.0
Far Below Average	36	24.0	24	15.4
ORF Rating				
Useful	128	85.3	112	71.8
Not Useful	22	14.7	44	28.2

Note. Third grade, N = 150. Fifth grade, N = 156.

Table 4

Correlations Between Predictors and Outcome

Variable		ORF Standard	CELDT	CST-ELA-RC
ORF Standard				
	Third	1.00		
	Fifth	1.00		
CELDT				
	Third	0.59*	1.00	
	Fifth	0.59*	1.00	
CST-ELA-RC				
	Third	0.47*	0.51*	1.00
	Fifth	0.52*	0.48*	1.00

Note. Third grade, N = 181. Fifth grade, N = 175. ORF Standard = Standardized oral reading fluency scores as measured by AIMSweb Reading-Curriculum Based Measurement. CELDT = California English Language Development Test. CST-ELA-RC = Standardized California Standards Test- English Language Arts- Reading comprehension scores. * $p < .01$.

Table 5

Third Grade Predictors of CST-ELA-RC Performance

	Interaction Model		Non-Interaction Model	
Variable	ß	Partial	ß	Partial
ORF Standard	.28	.26*	.26	.25*
CELDT	.37	.34*	.36	.33*
ORF Standard *CELDT	.09	-.07		
R^2	.32		.31	
F	27.07*		39.56*	
ΔR^2			-.007	
ΔF			1.71	

Note. N = 181. ORF Standard = Standardized oral reading fluency scores as measured by AIMSweb Reading-Curriculum Based Measurement; CELDT = California English Language Development Test; Partial = partial correlation coefficient. * $p < .01$.

Table 6

Fifth Grade Predictors of CST-ELA-RC Performance

Variable	ß	Partial
ORF Standard	.38	.35**
CELDT	.31	.28**
ORF Standard*CELDT	.15	.17*
R^2	.34	
F	28.68**	

Note. N = 175. ORF Standard = Standardized oral reading fluency scores as measured by AIMSweb Reading-Curriculum Based Measurement; CELDT = California English Language Development Test. * $p < .05$, ** $p < .01$.

Table 7

Summary of Research-Identified Word Callers

CELDT Level	Total N	Criterion 1 Proportion	Criterion 1 N	Criterion 2 Proportion	Criterion 2 N
Beginning/ Early Intermediate					
Third	71	1.5%	1	1.5%	1
Fifth	28	7.0%	2	0.0%	0
Intermediate					
Third	88	11.3%	10	2.2%	2
Fifth	84	9.5%	8	2.4%	2
Early Advanced/ Advanced					
Third	22	0.0%	0	0.0%	0
Fifth	63	7.9%	5	0.0%	0
Overall					
Third	181	6.0%	11	2.0%	3
Fifth	175	8.0%	15	1.1%	2

Note. CELDT = California English Language Development Test; Criterion 1 and 2 = word callers.

Table 8

Descriptive Statistics for Criterion 1 Word Callers

Variable	M	SD	Range
CELDT			
Third	473.18	13.08	448-499
Fifth	527.80	30.80	466-571
ORF Raw			
Third	117.64	8.82	108-141
Fifth	147.60	12.94	134-171
CST-ELA-RC			
Third	77.27	6.34	63.71-81.78
Fifth	77.35	6.70	63.93-84.20

Note. Third grade, N = 11. Fifth grade, N = 15. CELDT = California English Language Development Test. ORF Raw = Raw oral reading fluency scores as measured by AIMSweb Reading-Curriculum Based Measurement. CST-ELA-RC = Standardized California Standards Test- English Language Arts- Reading comprehension scores.

Table 9

Summary of ORF and CST-ELA-RC Difference Scores for Criterion 1 Word Callers

	Third Grade		Fifth Grade	
Standard Score Difference	Frequency	Percent	Frequency	Percent
10 – 15	2	18.2	3	20.1
16 – 20	3	27.3	3	20.1
21 – 25	3	27.3	3	20.1
26 – 30	3	27.3	6	40.2
Total	11		15	

Table 10

Third Grade Teacher-Nominated and Research-Identified Word Caller Comparisons

	Teacher-Nominated			
Research-Identified	Word Caller	Non-Word Caller	Indices	Pearson Chi-Square
	CELDT: Overall Sample, *N* = 150			
Word Caller	VP = 1	FN = 8	Sensitivity = 11%	
Non-Word Caller	FP = 16	VN = 125	Specificity = 89%	
	Pos. PV = 6%	Neg. PV = 94%	Hit Rate = 84%	χ^2= .000
	CELDT: Beginning and Early Intermediate, *N* = 60			
Word Caller	VP = 0	FN = 1	Sensitivity = 0%	
Non-Word Caller	FP = 7	VN = 52	Specificity = 88%	
	Pos. PV = 0%	Neg. PV = 98%	Hit Rate = 87%	χ^2= .134
	CELDT: Intermediate, *N* = 70			
Word Caller	VP = 1	FN = 7	Sensitivity = 13%	
Non-Word Caller	FP = 9	VN = 53	Specificity = 85%	
	Pos. PV = 10%	Neg. PV = 88%	Hit Rate = 77%	χ^2= .024
	CELDT: Early Advanced/ Advanced, *N* = 20			
Word Caller	VP = 0	FN = 0	Sensitivity = NA	
Non-Word Caller	FP = 0	VN = 20	Specificity = 100%	
	Pos. PV = NA	Neg. PV = 100%	Hit Rate = 100%	NA

Note. CELDT = California English Language Development Test; VP = valid positive; FN = false negative; FP = false positive; VN = valid negative; Pos. PV = positive predictive value [VP / (VP + FP)]; Neg. PV = negative predictive value [VN / (VN + FN)]; Sensitivity = VP / (VP + FN); Specificity = VN / VN + FP; Hit Rate = (VP + VN) / (VP + FN + VN + FP).

Table 11

Fifth Grade Teacher-Nominated and Research-Identified Word Caller Comparisons

	Teacher-Nominated			
Research-Identified	Word Caller	Non-Word Caller	Indices	Pearson Chi-Square
	CELDT: Overall Sample, *N* = 156			
Word Caller	VP = 2	FN = 13	Sensitivity = 13%	
Non-Word Caller	FP = 14	VN = 127	Specificity = 90%	
	Pos. PV = 13%	Neg. PV = 91%	Hit Rate = 83%	χ^2= .171
	CELDT: Beginning and Early Intermediate, *N* = 18			
Word Caller	VP = 1	FN = 1	Sensitivity = 50%	
Non-Word Caller	FP = 2	VN = 14	Specificity = 88%	
	Pos. PV = 33%	Neg. PV = 93%	Hit Rate = 83%	χ^2= 1.8
	CELDT: Intermediate, *N* = 79			
Word Caller	VP = 0	FN = 8	Sensitivity = 0%	
Non-Word Caller	FP = 6	VN = 65	Specificity = 92%	
	Pos. PV = 0%	Neg. PV = 89%	Hit Rate = 82%	χ^2= .732
	CELDT: Early Advanced/ Advanced, *N* = 59			
Word Caller	VP = 1	FN = 4	Sensitivity = 20%	
Non-Word Caller	FP = 6	VN = 48	Specificity = 89%	
	Pos. PV = 14%	Neg. PV = 92%	Hit Rate = 83%	χ^2= .346

Note. CELDT = California English Language Development Test; VP = valid positive; FN = false negative; FP = false positive; VN = valid negative; Pos. PV = positive predictive value [VP / (VP + FP)]; Neg. PV = negative predictive value [VN / (VN + FN)]; Sensitivity = VP / (VP + FN); Specificity = VN / VN + FP; Hit Rate = (VP + VN) / (VP + FN + VN + FP).

Table 12

ORF and CST-ELA-RC Performance by Word Caller Source

	Teacher-Nominated			Research-Identified		
Variable	*M*	*SD*	n	*M*	*SD*	n
ORF Standard						
Third	91.19	12.76	16	99.21	3.67	8
Fifth	91.67	8.85	14	100.87**	4.38	13
CST-ELA-RC						
Third	94.67	12.67	16	76.18**	7.09	8
Fifth	87.23	16.22	14	76.95*	6.94	13

Note. ORF Standard = Standardized oral reading fluency scores as measured by AIMSweb Reading-Curriculum Based Measurement. CST-ELA-RC = Standardized California Standards Test- English Language Arts- Reading comprehension scores. *$p < .05$. **$p < .01$.

Table 13

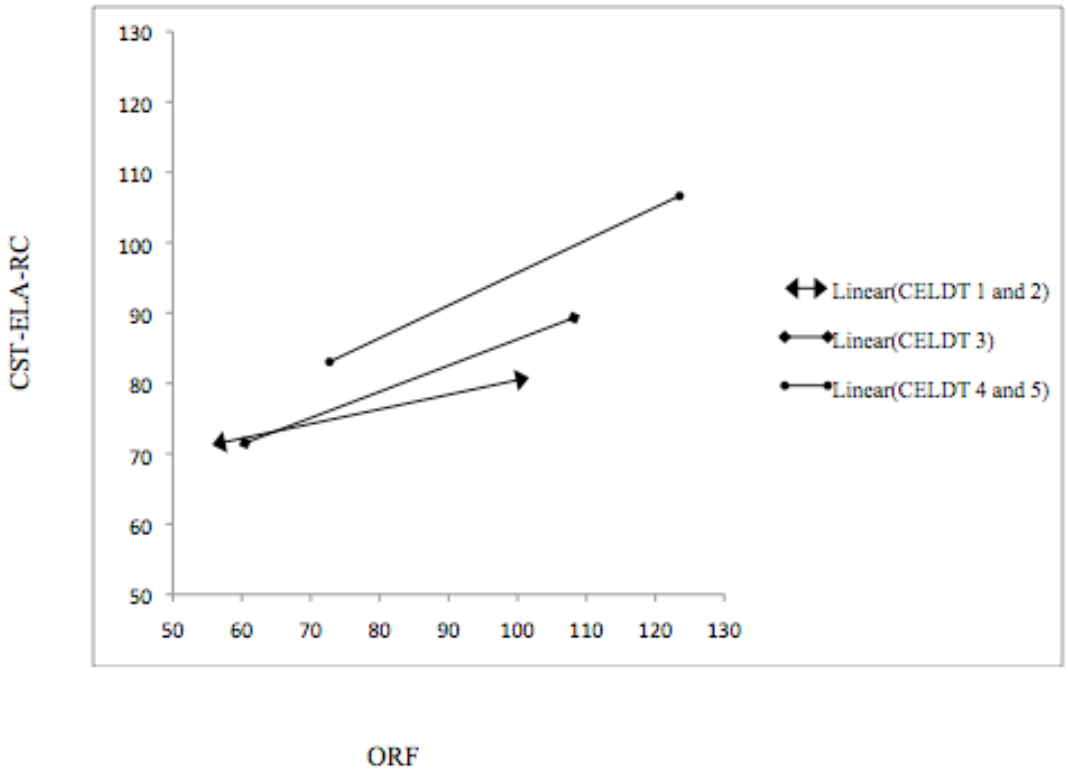
ORF and CST-ELA-RC Performance for Teacher-Nominated Categories

	Poor Readers[a]			Word Callers[b]			Comprehension Proficient[c]		
Variable	*M*	*SD*	n	*M*	*SD*	n	*M*	*SD*	n
ORF Standard									
Third	77.51[bc]	11.00	61	91.67[ac]	12.51	17	99.28[ab]	11.03	72
Fifth	83.27[bc]	12.75	70	92.20[a]	8.37	16	92.42[a]	10.58	70
CST-ELA-RC									
Third	82.77[bc]	13.88	61	93.91[a]	12.67	17	97.68[a]	13.29	72
Fifth	82.14[c]	11.22	70	86.32	15.38	16	88.56[a]	13.70	70

Note. ORF Standard = Standardized oral reading fluency scores as measured by AIMSweb Reading-Curriculum Based Measurement. CST-ELA-RC = Standardized California Standards Test- English Language Arts- Reading comprehension scores. Superscript indicates significantly different scores with $p < .05$.

Figure 1

ORF and CST-ELA-RC Relationship for Fifth Grade Stratified by CELDT Levels



Appendix A

Teacher Survey

Please complete items 1-9, place this survey in the envelope labeled "Teacher Survey", seal the envelope and return it to the research box in the office. Thank you!

1. Gender:
☐ **Male**
☐ **Female**

2. Age: _____

3. Ethnicity:
☐ **American Indian or Alaska Native**
☐ **Asian**
☐ **Hispanic or Latino**
☐ **Black or African American**
☐ **White, Non-Hispanic**
☐ **Native Hawaiian or Pacific Islander**

4. Type of Credential/ Certification: _____

5. Year that your most recent credential/certification was completed: _____

6. Highest Degree Obtained:
☐ **BA/BS**
☐ **MA/MS**
☐ **PhD**
☐ **PsyD**
☐ **Other:_____**

7. How many years have you been teaching and/or holding a job as a certificated staff member? _____

8. What is your current grade assignment?
☐ **Grade 3**
☐ **Grade 5**

9. Do you hold certification to teach English learners?
☐ **Yes**
☐ **No**

Appendix B

Classroom Survey

After completion, please place this survey in the envelope labeled "Classroom Survey", seal the envelope and return it to the research box in the office. Thank you!

1. Please fill in the following information regarding each of your students listed below. Consider each student separately from the others when choosing your responses.
    a. Estimate the skill level for each student for the following categories: *reading fluency* and *reading comprehension.* When considering skill level use grade level expectations.

    b. Indicate your response to the question:
    Is oral reading fluency (ORF) a useful assessment to <u>screen this particular student for reading problems and determine if intervention is needed</u>?

    ORF is defined as a task in which a student reads from a grade level appropriate passage of connected, meaningful text for a discrete time period. The examiner records the number of words read correctly in this time period. Omissions, mispronunciations, substitutions, and hesitations longer than three seconds are counted as incorrect. The total score is the number of words read correctly in one minute.

Student Name	**Reading Fluency Level (select one)**	**Reading Comprehension Level (select one)**	**Answer to question 1b (above)**
Last, First	☐ Above average ☐ Average ☐ Below average ☐ Far below average	☐ Above average ☐ Average ☐ Below average ☐ Far below average	☐ Yes ☐ No
Last, First	☐ Above average ☐ Average ☐ Below average ☐ Far below average	☐ Above average ☐ Average ☐ Below average ☐ Far below average	☐ Yes ☐ No
Last, First	☐ Above average ☐ Average ☐ Below average ☐ Far below average	☐ Above average ☐ Average ☐ Below average ☐ Far below average	☐ Yes ☐ No
Last, First	☐ Above average ☐ Average ☐ Below average ☐ Far below average	☐ Above average ☐ Average ☐ Below average ☐ Far below average	☐ Yes ☐ No
Last, First	☐ Above average ☐ Average ☐ Below average ☐ Far below average	☐ Above average ☐ Average ☐ Below average ☐ Far below average	☐ Yes ☐ No