Screening for At-Risk Readers: Use of Curriculum-Based Measurement and Teacher Judgment

https://escholarship.org/uc/item/24k5r5d1

Tung, Catherine Yang

2013

Peer reviewed|Thesis/dissertation
Screening for At-Risk Readers: Use of Curriculum-Based Measurement and Teacher Judgment

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

Doctor of Philosophy

in

Education

by

Catherine Yang Tung

June 2013

Dissertation Committee:

Dr. Mike Vanderwood, Chairperson
Dr. Gregory Palardy
Dr. Cathleen Geraghty Jenkinson
The Dissertation of Catherine Yang Tung is approved:

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

Committee Chairperson

University of California, Riverside
ABSTRACT OF THE DISSERTATION

Screening for At-Risk Readers:
Use of Curriculum-Based Measurement and Teacher Judgment

by

Catherine Yang Tung

Doctor of Philosophy, Graduate Program in Education
University of California, Riverside, June 2013
Dr. Mike Vanderwood, Chairperson

This study examined the predictive validity and accuracy of using teacher judgment with curriculum-based measurement as a part of screening for at-risk readers. The participants included native English speakers and English language learners in third and fifth grade. Results from the predictive validity analyses indicated that after controlling for prior year’s test performance and curriculum-based measurement (e.g.: R-CBM and Maze), teacher judgment contributed unique variance in predicting future reading performance. Results from the predictive accuracy analyses indicated that established cut-scores for curriculum-based measures yielded relatively lower sensitivity and higher specificity. Regarding the effect of student variables, it was found that English language proficiency and level of reading achievement moderated the contribution of teacher judgment in predicting future reading performance. However, teachers were not found to be a source of variability in the accuracy of their judgments.
Table of Contents

Introduction ................................................................................................................... 1
Method ......................................................................................................................... 48
Results ......................................................................................................................... 54
Discussion ..................................................................................................................... 68
References .................................................................................................................... 87
Appendix A .................................................................................................................... 101
List of Tables

Table 1: Descriptive Statistics..........................................................93
Table 2: Correlations.................................................................94
Table 3: Presence of Significant Differences Between Correlations.................95
Table 4: Research Question 2a Proportion of Variance Explained.....................96
Table 5: Research Question 2b Proportion of Variance Explained.....................97
Table 6: Sensitivity, Specificity, and AUC Values.......................................98
Table 7: Cut Scores and Specificity Values............................................99
Table 8: Research Question 4 Proportion of Variance Explained...................100
Screening for At-Risk Readers:
Use of Curriculum-Based Measurement and Teacher Judgment

Reading is a crucial skill that individuals need in order to act as functional members in society. Deficient reading skills have been found to be associated with negative educational and social outcomes (McGill-Franzen, 1987). Despite this, many students today experience difficulties with reading. According to the National Center for Education Statistics (2011), over half of fourth-grade students in 2009 scored below the Proficient level on the reading portion of the National Assessment of Educational Progress (NAEP). This indicates that a significant number of students do not possess adequate reading skills. A growing number of today’s students are from ethnic minority backgrounds, many of whom are also English language learners (National Center for Education Statistics, 2011). Over four million ELLs reside in the United States, a majority of whom speak Spanish as their native language (Kindler, 2002). As a whole, ELL students start significantly behind their non-ELL peers upon school entry (Rumberger & Gandara, 2004). According to a survey of 41 states, less than 20% of ELLs met state-established criteria for reading comprehension (Kindler, 2002). In light of dim statistics regarding the reading skills of today’s students, including ELLs, it is important to have adequate screening measures in order to identify these at-risk students.

Characteristics of Screening Measures

When selecting an approach to screening, it is important to consider the screeners’ intended use (Glover & Albers, 2007; Jenkins, Hudson, & Johnson, 2007). The appropriateness of using a screener depends on its theoretical support, empirical support,
degree of alignment with target constructs, and compatibility with the needs surrounding local service delivery (Glover & Albers, 2007). An ideal screener should have adequate reliability, validity, and efficiency (Glover & Albers, 2007; Jenkins et al., 2007).

**Validity and classification accuracy.** As with all assessments, the evaluation of a screener should include an examination of various domains of validity (e.g.: content, construct, criterion, concurrent, predictive, consequential) (Glover & Albers, 2007; Jenkins et al., 2007). However, the following discussion will focus on issues of predictive validity, which is one of the major areas of focus for studies involving screening. The case for predictive validity is typically established by examining the correlation between a target screener and other established measures of the same construct (Jenkins et al., 2007). After conducting correlational analyses, it is important to consider classification accuracy by examining values of sensitivity, specificity, positive predictive value, negative predictive value, and overall hit rate (Glover & Albers, 2007; Jenkins et al., 2007). Sensitivity is the proportion of students with poor later outcomes who were correctly identified as at-risk by the screener (Glover & Albers, 2007). Specificity is the proportion of students with positive later outcomes who were correctly identified as not at-risk by the screener. Positive predictive value is the proportion of correctly identified students out of all who were identified as at-risk by the screener. Negative predictive value is the proportion of correctly identified students out of all who were identified as not at-risk by the screener. Overall hit rate is the proportion of all students who were correctly identified. The following are the calculations used for the classification accuracy values: sensitivity (valid positives / [valid positives + false negatives]),
specificity (valid negatives / [valid negatives + false positives]), positive predictive value (valid positives / [valid positives + false positives]), and negative predictive value (valid negatives / [valid negatives / false negatives]) (Glover & Albers, 2007). Factors that affect the classification accuracy of screeners include the nature of the criterion measure, degree of match between the constructs measured by the screener and the criterion measure, duration between time of screening and administration of the criterion measure, instruction or intervention that occurs between screening and administration of the criterion measure, and cut-score used for classification (Jenkins et al., 2007).

Efficiency and usability. In addition to examining the validity of screening instruments, it is also important to consider issues of efficiency and usability (Glover & Albers, 2007; Jenkins et al., 2007). This may be done by carefully analyzing the costs and benefits associated with administering the screener (Jenkins et al., 2007). Other areas to examine include acceptability, administration feasibility, and usefulness of outcomes (Glover & Albers, 2007). The idea of having useful outcomes it tied to the concept of consequential validity. According to Messick (1994; 1995), it is not only important to examine score meaning (construct validity), but also score use (consequential validity). Consequential validity is examined by considering the intended and unintended consequences of the screening measure.

Reading Theory

Various models of reading have been posited in the literature. Hoover and Gough’s (1990) simple view of reading asserts that both decoding and listening comprehension influence reading comprehension. This relationship can be
conceptualized as “reading comprehension = decoding x listening comprehension.”

Therefore, the more fluent readers are at decoding, the more resources become available for comprehension. According to LaBerge and Samuels’ (1974) automaticity model and Posner & Snyder’s (1975) interactive model of reading, reading requires the coordination of many processes during a short amount of time. Automaticity plays a crucial role in reading competence. Skilled readers devote less attention to the lower skills of reading (e.g.: decoding and word identification) and more attention to the higher skills (e.g.: comprehension). This same concept is also present in Perfetti’s (1985) verbal efficiency theory. According to this theory, good readers decode faster than poor readers, which leaves more attentional resources to devote to comprehension. In considering these above theories, it can be seen that reading fluency can act as an indicator of reading comprehension (Fuchs, Fuchs, Hosp, & Jenkins, 2001).

According to theory and research, oral reading fluency has been considered to be an indicator of overall reading performance (Fuchs et al., 2001). Components of oral reading fluency include both reading speed and accuracy (Fuchs et al., 2001). Research suggests that reading fluency is highly related to comprehension up to sixth grade, after which vocabulary starts to have more of an influence (Perfetti, 1985). The greatest growth in oral reading fluency happens in the early elementary grades with growth slowing as students become older (Fuchs et al., 2001).

**Measurement of Reading Using ORF and Maze**

Various studies support the use of oral reading fluency (ORF) measures to assess general reading performance (Fuchs et al., 2001). These measures assess student
performance in the oral reading of connected text. Compared with other measures of fluency, such as isolated word reading fluency and silent reading fluency, ORF has been found to exhibit a stronger relationship with reading comprehension (Fuchs et al., 2001). The following is a discussion of ORF in comparison to measures of reading comprehension (e.g.: Maze).

A study by Silberglitt, Burns, Madyn, and Lail (2006) compared ORF and Maze in their relationships with high-stakes standardized assessments. The participants included 5,427 third, fifth, seventh, and eighth grade students. The sample was 94% Caucasian. ORF was administered to each grade with the median of three probes taken as the final score. Maze was administered to seventh and eighth graders. The outcome measure for third, fifth, and seventh graders was the Minnesota Comprehensive Assessments – Reading (MCA-R). The outcome measure for eighth graders was the Basic Standards Test – Reading (BST-R). Both outcome measures were administered two months after ORF and/or Maze. Correlations between ORF and the outcome measures were .71 for third graders, .65 for fifth graders, .60 for seventh graders, and .51 for eighth graders. These correlations appeared to decrease as grade increased. Comparisons using the Fisher transformation found the correlation for third graders to be significantly larger than the correlations for all other grades. Also, the correlation for fifth graders was significantly larger than the one for eighth graders. Correlations between Maze and the outcome measures were .54 for seventh graders and .49 for eighth graders. Although correlations between ORF and the outcomes measures appear to be larger than the correlations between Maze and the outcome measures, no significant difference was
found between these correlations. This study provides evidence for the relationship of both ORF and Maze with high-stakes standardized assessments.

Graney, Martinez, Missal, and Aricak (2010) compared the technical adequacy of ORF and Maze for screening in the late elementary grades. Both reliability (e.g.: test-retest and alternate form) and validity (e.g. criterion and predictive) were examined. The participants included 76 students who were followed from fourth to fifth grade. The sample was 93% Caucasian. ORF and Maze from AIMSweb were administered during the fall, winter, and spring of both grades. For ORF, the median of three probes taken as the final score. Alternate form and test-retest reliability probes were given two weeks following the spring administration in fourth grade. The Test of Reading Comprehension – Third Edition (TORC-3) paragraph reading subtest was administered in the spring of fourth grade. The Indiana Statewide Testing for Educational Progress – Plus (ISTEP+) was administered in the fall of fifth grade with the English/language arts section used as the outcome. Pearson correlations were used to examine reliability and criterion validity. ORF yielded .96 test-retest reliability, .91 alternate form reliability, .31 criterion validity with the TORC-C, and .72 criterion validity with the ISTEP+. Maze yielded .89 test-retest reliability, .82 alternate form reliability, .41 criterion validity with the TORC-C, and .67 criterion validity with the ISTEP+. Path analysis was used to examine predictive validity, which yielded a .31 path coefficient between spring of fourth grade ORF and ISTEP+ and a .41 path coefficient between spring of fourth grade Maze and ISTEP+. Based on these results, the authors concluded that both ORF and Maze could serve as appropriate screeners in fourth and fifth grade.
In another study, Fuchs, Fuchs, and Maxwell (1988) compared the criterion validity of ORF with other direct measures of reading comprehension. The participants included 70 middle school students with reading disabilities. Students were administered ORF in addition to comprehension measures including passage recall, cloze, and question answering. The Stanford Achievement Test (SAT-10) reading comprehension subtest was used as the outcome measure. The results found that ORF had a significantly higher correlation (.91) with the SAT-10 than did all the other direct measures of comprehension (.70 for passage recall, .72 for cloze, and .82 for question answering).

**Predictive Validity of ORF**

In a meta-analysis, Reschly, Busch, Betts, Deno, and Long (2009) reviewed the correlational evidence between ORF and standardized measures of reading performance. The analysis included technical reports and journal articles studying the relationship between ORF and some other outcome measure in reading with first through sixth graders. The authors chose to exclude dissertations, conference proceedings, studies using achievement scores to predict ORF, studies combining data across grades, and studies that did not use standardized procedures for ORF. Potential moderating variables examined in the analysis include grade, outcome test administration format (e.g.: group vs. individually administered), type of reading score used in the outcome measure (e.g.: decoding, word identification, vocabulary, comprehension, or total reading), type of outcome measure (state assessments vs. national norm-referenced tests), and length of time between administration of ORF and the outcome measure. Hierarchical linear modeling was used to analyze 289 correlations coefficients nested within studies.
Overall, a strong correlation was found between ORF and the reading outcome measures. The median coefficient was .68 and the average weighted coefficient was .67. Correlations across grades were not significantly different. However, higher correlations were found for individually administered tests ($r = .83$) than for group-administered tests ($r = .71$). Higher correlations were found between ORF and nationally norm-referenced tests ($r = .74$) than between ORF and state assessments ($r = .65$). The authors suggested that this difference might be due to the relatively high technical quality of national tests. Regarding the type of reading score used in the outcome measure, ORF exhibited a significant relationship with decoding, vocabulary, and comprehension measures. However, the relationship between ORF and word identification measure was significantly stronger. Lastly, it was found that correlations decreased as time duration increased between the administration of ORF and the outcome measure. The authors concluded that ORF served as an adequate and efficient indicator of future reading performance considering the overall positive results and relatively low costs associated with using ORF. Due to the variability among the studies, correlations were not examined as a function of student variables such as socioeconomic status, ethnicity, or English language learner status. Also, it is important to note that the lack of significant differences across grades in this study is in contrast with the findings of other studies. In addition to this meta-analysis, numerous individual studies have examined the relationship between ORF and later reading outcomes as measured by standardized assessments. The following is a review of some of these studies.
Studies not controlling for previous year’s test. Goffreda, Diperna, and Pedersen (2009) examined the predictive validity of various DIBELS measures, including ORF. The participants of this study, 67 first graders, were administered DIBELS Letter Naming Fluency (LNF), Phoneme Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), and ORF during the winter. The TerraNova California Achievement Test (CAT) was given in second grade, and the Pennsylvania System of School Assessment (PSSA) was given in third grade. Logistic regression and ROC analyses were used to assess the predictive validity of the screening measures. Among the various measures, only ORF was found to be a significant predictor of later reading outcomes. Using the DIBELS cut-scores for risk, ORF yielded 80% sensitivity and 87% specificity for predicting second grade performance on the CAT. ORF yielded 77% sensitivity and 88% specificity for predicting third grade performance on the PSSA. By adjusting the DIBELS cut-scores, predictive accuracy for the PSSA was increased to 88% sensitivity and specificity.

McGlinchey and Hixson (2004) examined the validity of ORF for predicting reading outcomes. The participants of this study included 1,362 fourth graders from a period of eight years. The sample was about 50% Caucasian, 45% Black, and 60% from low socioeconomic backgrounds. ORF from a district basal text was administered two weeks before the Michigan Education Assessment Program (MEAP) reading assessment, which measured mostly reading comprehension. Correlations between ORF and MEAP ranged from .49 to .81. Based on previous research, the authors selected 100 words correct per minute to use as the cut-score for ORF in predicting MEAP performance. This yielded a sensitivity of .75, specificity of .74, positive predictive power of .77, negative
predictive power of .72, and overall correct classification of .74. The positive and negative predictive power values were found to be an improvement over a .54 base rate of failing and .46 base rate of passing the MEAP. Cohen’s kappa, which corrects for chance predictions, was found to be .48. Although this study purports to examine prediction, it is important to note that ORF was given only two weeks before the MEAP.

In a longitudinal study, Hintze and Silberglitt (2005) examined the predictive validity of ORF with a sample of five cohorts of 1,766 students followed from first to third grade. The participants were 94% Caucasian, 5% with special needs, and 30% from low socioeconomic backgrounds. ORF was administered in fall, winter, and spring starting in the winter of first grade through the spring of third grade. The Minnesota Comprehensive Assessment (MCA), administered at the end of third grade, served as the reading outcome measure. The results indicated that ORF performance at each grade was significantly related to MCA performance with correlations ranging from .49 to .69. In addition to correlation analyses, the authors also compared various approaches (e.g.: logistic regression, discriminant analysis, and ROC analysis) for creating ORF cut-scores to predict MCA outcomes. It was found that cut-scores from each approach yielded adequate diagnostic accuracy values. The range of diagnostic accuracy values across approaches were as follows: .50 to .79 sensitivity, .76 to .92 specificity, .58 to .86 positive predictive power, and .52 to .88 negative predictive power. Generally, specificity and negative predictive values of the cut-scores were higher than sensitivity and positive predictive values.
**Studies controlling for previous year’s test.** Although most studies examining the validity of ORF in predicting performance on high-stakes assessments do not include prior year’s high-stakes test performance as a predictor, there is evidence in the literature that ORF makes a significant predictive contribution independent of previous year’s performance on high-stakes assessments (Baker et al., 2008; Johnson, Jenkins, & Petscher, 2010; Wood et al., 2006). Baker et al. (2008) examined the predictive validity of ORF for performance on high-stakes reading assessments above and beyond the predictive contribution of previous year’s high-stakes assessment performance.

Participants included over 9,600 first through third graders from Reading First schools. DIBELS ORF was administered during fall, winter, and spring. The Stanford Achievement Test (SAT-10) was given in the spring of first and second grade. The Oregon Statewide Reading Assessment (OSRA) was given in the spring of third grade. Prediction models were used to examine the contribution of ORF and previous year’s high-stakes test score in predicting the current year’s high-stakes test score. When predicting second grade performance on the SAT-10, one prediction model found that ORF contributed 70% of the variance. Another prediction model found that ORF and first grade SAT-10 combined contributed 76% of the variance. When predicting third grade performance on the OSRA, one prediction model found that ORF contributed 52% of the variance. Another prediction model found that ORF and second grade SAT-10 combined contributed 59% of the variance. This study supports the use of ORF as a screener because it demonstrated that ORF still made a significant contribution to the prediction even with previous years’ high-stakes assessment included as a predictor.
Wood et al. (2006) examined the predictive validity of ORF for performance on the Colorado Student Assessment Program (CSAP) reading test controlling for the contribution of previous year’s CSAP performance. The participants were 281 third through fifth graders native English speakers from one elementary school. DIBELS ORF was administered in the winter, and the CSAP was administered two months later. Correlation analyses yield values .67 to .75 between ORF and CSAP. Multiple regression was used to assess the predictive contribution of ORF above previous year’s CSAP performance. In predicting both fourth and fifth grade CSAP performance, the results of simultaneous multiple regression showed that both previous grade’s CSAP performance and current grade’s ORF were significant and independent predictors. When previous grade’s CSAP performance was entered as the first predictor and current grade’s ORF as the second predictor, ORF contributed an additional 9% of the variance in predicting CSAP performance in fourth grade and an additional 4% of the variance in fifth grade. In predicting CSAP performance, ORF yielded high sensitivity (.86 to .95) and lower specificity (.58 to .67) values across grades.

In addition to the previous two studies, Johnson et al. (2010) also considered the use of previous year’s high-stakes tests scores in predicting performance in the current year. The outcome measure was third grade performance on the Florida Comprehensive Assessment Test – Sunshine State Standards (FCAT-SSS). The predictors were second and third grade DIBELS ORF, second grade PPVT, and second grade SAT-10. The participants included over 12,000 students. Although the focus of this study was the effect of combining various screening measures in order to maximize accuracy in
predicting third grade reading performance, it is noted that the resulting “best” combination of predictors included both ORF and second grade SAT-10 in addition to other measures. Therefore, this study supports the use of ORF in predicting performance on high-stakes assessments even when previous year’s high-stakes test score is included among the predictors. The above three studies are consistent in supporting the predictive validity of ORF for performance on high-stakes reading assessments even when considering the predictive value of prior year’s high-stakes test performance (Baker et al., 2008; Johnson et al., 2010; Wood et al., 2006).

Sources of Bias for ORF

Bias in testing has been conceptualized in various ways. One simple conceptualization of test fairness requires the absence of differences in mean scores across groups (Linn, 1973). This definition of fairness assumes that different groups do not vary on the variable being assessed. It also assumes that any mean differences across groups are due to bias within a test. However, this is definition of test bias is inadequate because oftentimes different groups in a population do indeed vary on a given variable (Linn, 1973). According to Linn (1973) and Messick (1995), the validity of a test is assessed based on its meaning and intended use. Therefore, a test can be either biased or unbiased depending on the purpose of testing and interpretation of the results. Test bias occurs when a given test score is differentially valid for any subgroup of individuals who have taken the test (Cole & Moss, 1993; Linn, 1973). According to Cole and Moss (1993), test bias may originate from the following sources: 1) presence of test administration factors that limit the performance of certain groups, 2) presence of content
bias within the test (e.g.: cultural-specific knowledge, stereotypes, and lack of diversity), 3) presence of construct irrelevant variance within the test for its intended use, and 4) presence of differential criterion-related validity across groups of test takers. When a test is biased, the interpretation of a given test score differs across groups of individuals (AERA/APA/NCME, 1999; Cole & Moss, 1993; Linn, 1973). This may occur across various types of subgroups such as ethnicity, gender, socioeconomic status, special needs status, and language status.

Although there are different ways of conceptualizing bias, the focus of bias in the screening literature is on predictive bias. According to Linn (1973), it is important to focus on predictive validity as a possible source of bias when the purpose of a test is to predict outcomes. According to educational and psychological testing standards as set forth by AERA/APA/NCME (1999), as well as other many other psychometricians (Cole & Moss, 1993; Linn, 1976), predictive bias occurs when a test makes more accurate predictions for one group than it does for another group. For example, in the context of screening, this occurs when a given score on a screener predicts positive later outcomes for one group but negative later outcomes for another group. One common method of assessing for predictive bias is through comparing regression equations among different subgroups of test takers (Jenson, 1980; Linn, 1976). The following is a discussion on possible sources of predictive bias in ORF.

**Ethnicity.** Various studies have explored ethnicity as a possible source of bias for using ORF to predict performance on high-stakes standardized tests (Hintze, Callahan, Matthews, Williams, & Tobin, 2002; Hixson & McGlinchey, 2004; Kranzler, Miller, &
Jordan, 1999; Pearce & Gayle, 2009). Kranzler et al. (1999) examined the validity of ORF with a sample of 326 Caucasian and Black second through fifth graders. ORF passages from the school’s reading textbook were administered in March. The reading comprehension portion of the California Achievement Test (CAT) was administered shortly afterwards in the spring. Correlations between ORF and the CAT ranged from .51 to .63. Simultaneous multiple regression was conducted for each grade with the predictors including ORF score, gender, and ethnicity. No bias was found in second and third grade. However, intercept bias was found in fourth and fifth grade. Specifically, ORF overestimated reading comprehension for the Black students and underestimated for the Caucasian students. From this study, grade appears to be a moderator of the presence of ethnicity bias. In the discussion, the authors state that the presence of bias should not mean the discontinued use of ORF. Instead, the authors suggest that different cut-scores may be needed for different groups in order to maximize predictive validity.

Similar to the previous study, Hintze et al. (2002) examined the validity of ORF with a sample of 136 Caucasian and Black second through fifth graders. Third grade ORF passages and the reading comprehension subtest from the Woodcock Johnson Psychoeducational Battery – Revised (WJ-R) were administered during the same day. The results from hierarchical multiple regression indicated that ethnicity did not affect prediction of reading comprehension scores after controlling for students’ age, gender, and socioeconomic status. However, a second regression analysis separating the ethnic groups found a possible bias by ethnic group. It was found that age and ORF contributed 58% of the variance in reading comprehension for Black students and 30% of the
variance for Caucasian students. Although it was interesting that age and ORF contributed much more of the variance in outcomes for Blacks than it did for Caucasians, the authors concluded that ORF generally does have bias. It is important to note some characteristics of this study that may limit the generalizability of its findings. First, ORF and the outcome measure were administered during the same day. Thus, this study does not examine predictive validity. Second, regardless of their grades, all students were administered third grade ORF passages.

Hixson and McGlinchey (2004) examined the predictive validity of ORF with a sample of 442 Caucasian and Black fourth graders. ORF from a district basal was administered two weeks before the Michigan Educational Assessment Program’s (MEAP) reading assessment and four months before the Metropolitan Achievement Tests – 7th Edition’s (MAT/7) total reading portion. Mixed results regarding bias by ethnicity was found depending on method of data analysis. Under simultaneous multiple regression, ORF, socioeconomic status, and ethnicity were all significant predictors on later reading outcomes on both MEAP and MAT/7. On both outcomes measures, scores of higher socioeconomic status and Caucasian students were overestimated. Scores of lower socioeconomic status and Black students were underestimated. Despite this, no evidence of bias by ethnicity or socioeconomic status was found when using stepwise multiple regression.

Pearce and Gayle (2009) examined the predictive validity of ORF with a sample of 543 Caucasian and Native American third graders. DIBELS ORF was given in the winter. The reading comprehension subtest of the Dakota State Test of Educational
Proficiency (DStep) given four months later was used as the outcome measure. The results from multiple regression indicated that ORF contributed 40% of the variance. Socioeconomic status contributed an additional 2%, and ethnicity contributed an additional 3%. Diagnostic accuracy analyses indicated that ORF yield high specificity and negative predictive power but lower sensitivity and positive predictive power. The prevalence of false negatives was found to be higher among the Native American students.

Overall, studies examining the predictive bias of ORF by ethnicity have yielded mixed results (Hintze et al., 2002; Hixson & McGlinchey, 2004; Kranzler et al., 1999; Pearce & Gayle, 2009). Evidence of possible bias was found in some studies. However, sometimes evidence for the presence or absence of bias was found within the same study with differing results depending on the data analysis method used (e.g.: Hintze et al., 2002; Hixson & McGlinchey, 2004). In addition, possible moderator variables such as grade and socioeconomic status make the results of these studies less conclusive.

Language status. There have been some studies in the literature that have examined language status as a possible source of bias for using ORF to predict performance on high-stakes standardized tests (Hosp, Hosp, & Dole, 2011; Klein & Jimerson, 2005; Muyskens, Betts, Lau, & Marston, 2009; Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008). Klein and Jimerson (2005) examined the effect of ethnicity, gender, socioeconomic status, and home language on the concurrent and predictive validity of ORF. The participants included about 4,000 Caucasian and Hispanic first through third graders. ORF was administered in the fall and spring. The total reading
portion of the Stanford Achievement Test – Ninth Edition (SAT-9) administered in the spring was used as the outcome measure. Hierarchical multiple regression was used to analyze the data. After controlling for ethnicity, differences in ORF and the SAT-9 scores were found as a function of socioeconomic status and home language. Specifically, Hispanic students from low socioeconomic backgrounds scored significantly lower than did Hispanic students from higher socioeconomic backgrounds. Also, Spanish-speaking Hispanic students scored significantly lower than did English-speaking Hispanic students. For the concurrent validity analyses, significant intercept bias was found between English-speaking Caucasian students and Spanish-speaking Hispanic students. For the predictive validity analyses, some evidence of slope bias was found between English-speaking Caucasian students and Spanish-speaking Hispanic students. Specifically, ORF underestimated reading performance for English-speaking Caucasian students and overestimated for Spanish-speaking Hispanic students. In the discussion, the authors concluded that home language and ethnicity interact to produce bias with neither factor alone producing bias. For example, no significant bias were found between Hispanic students who spoke Spanish and those who spoke English. Biases were also absent among English-speaking Hispanic and Caucasian students. However, the main factor contributing to bias was home language status. Mixed results were found regarding the effect of socioeconomic status.

Roehrig and colleagues (2008) examined the predictive validity of ORF as a function of ethnicity, socioeconomic status, and English language learner status. The participants included 35,207 third graders from Reading First schools. DIBELS ORF was
administered during four benchmark periods throughout the year. The Florida Comprehensive Assessment Test – Sunshine State Standards (FCAT-SSS) and the Stanford Achievement Test (SAT-10) administered in the spring were used as the outcome measures for reading comprehension. The results from logistic regression indicated that ORF predicted performance on the FCAT-SSS regardless of socioeconomic status, ethnicity, or English language learner status.

Similar to the previous studies, Hosp and colleagues (2011) examined the effect of ethnicity, disability status, socioeconomic status, and English language learner status on the predictive validity of ORF. The participants included 3,805 first through third graders from Reading First schools. DIBELS ORF was administered during the fall, winter, and spring from the winter of first grade to the spring of third grade. The English / Language Arts portion of the Utah State Criterion-Referenced Tests (UCRTs) administered in the spring was used as the outcome measure. ROC analyses were used to calculate curves for each group within each category (e.g.: ethnicity, disability status, socioeconomic status, and English language learner status). Sensitivity, specificity, and AUC were compared between groups within each category. A value above .80 was considered “adequate.” Regarding socioeconomic status, ORF yielded adequate AUCs and sensitivity, but low specificity for both higher and lower socioeconomic groups. For ethnicity, some significant differences in AUC, sensitivity, and specificity were found between groups in certain grades. For English language learner status, AUCs and sensitivity was generally adequate, but specificity was low for both groups. There were no significant differences between ELLs and native English speakers.
Muyskens and colleagues (2009) examined the concurrent and predictive validity of ORF with a sample of ELLs from different language backgrounds. The participants were 1,205 fifth grade ELLs whose native language was Spanish, Hmong, or Somali. ORF from a district basal was administered in the fall. The Minnesota Comprehensive Assessment (MCA) administered in the spring was used as the outcome measure. Regression analyses found ORF to be a significant predictor of MCA performance for all students. ROC analyses were conducted for each language group. ORF was found to predict MCA equally well for each language group as indicated by the presence of overlapping AUC confidence intervals. One difference to note between this study and the two above studies (e.g. Hosp et al., 2011; Klein & Jimerson, 2005) is that all participants in this study were ELLs. Comparisons were made between native language groups but were not made with native English speakers.

In addition to the above studies examining language status as a possible source of bias, other studies have also touched upon this issue (Johnson, Jenkins, Petscher, & Catts, 2009; Johnson et al., 2010; Wiley & Deno, 2005). Wiley and Deno (2005) compared the value of adding a comprehension measure to ORF in predicting reading performance on the Minnesota Comprehensive Assessment (MCA) for native English speakers and English language learners. It was found that Maze accounted for significant variance beyond ORF in predicting MCA performance only for native English speakers. A study by Johnson and colleagues (2009) examining the predictive validity of a variety of measures for performance on the SAT-10 in first grade found that the best predictors were the same for native English speakers and ELLs. However, it was also found that
lower ORF cut-scores were needed for ELLs than for native English speakers in order to obtain 90% sensitivity in prediction. In another study examining the predictive validity of a variety of measures for performance on the third grade reading performance on the Florida Comprehensive Assessment Test – Sunshine State Standards (FCAT-SSS), Johnson and colleagues (2010) found English language learner status to be included as one of the predictors among an optimal combination of factors predicting reading performance.

As a whole, the literature examining the predictive bias of ORF by English language learner status has yielded conflicting results. Some studies have found evidence for bias (e.g.: Klein & Jimerson, 2005; Johnson et al., 2009; Johnson et al., 2010; Wiley & Deno, 2005) while others studies have found the contrary (e.g.: Hosp et al., 2011; Muyskens et al., 2009; Roehrig et al., 2008). To date, it is difficult to make conclusive statements about the effect of English language learner status on the predictive validity of ORF. One reason for this is that studies have typically characterized ELLs as a homogeneous group or, at best, as separated by native language. English language proficiency has not been considered as a source of bias. However, it is common knowledge that ELLs come to school with a wide range of English language skills. Perhaps, future research examining English language proficiency will help shed more light on this topic.

**Improving Screening Accuracy of ORF**

**Considering growth data.** In exploring various methods of improving screening accuracy, researchers have examined the role of using ORF growth data in the prediction
of later reading outcomes as measured by high-stakes state assessments or public norm-referenced tests (Baker et al., 2008; Keller-Margulis, Shapiro, & Hintze, 2008; Schatschneider, Wagner, & Crawford, 2008; Stage & Jacobsen, 2001). In these studies, researchers have typically examined the predictive contribution of ORF growth data above and beyond the contribution of initial ORF level. At face value, one advantage of this approach is that it considers students’ response to instruction and relies on more than one data point to predict future performance. However, a disadvantage of using growth data is that it requires a longer period of time before a prediction is made about students’ future performance, thereby delaying possible needed intervention services (Jenkins et al., 2007).

Schatschneider and colleagues (2008) compared the validity of using ORF level, ORF growth, and ORF level + growth to predict reading outcomes as measured by the Stanford Achievement Test (SAT-10). The participants included 23,438 first graders from Reading First schools followed into second grade. Of the participants, 72% were from low socioeconomic backgrounds, 41% were Caucasian, 32% were Black, and 22% were Hispanic. DIBELS ORF was given in September, December, February, and April of first grade as well as in September of second grade. The median of three passages was taken as the score. The SAT-10 was given in the spring of both first and second grade. Growth curve analyses were used to estimate ORF growth at the four measurement time points in first grade. Both a linear and a quadratic growth model were examined. It was found that the linear model produced more reliable growth estimates, but the quadratic model was a better fit with the data. Following these analyses, two hierarchical multiple
regression analyses were conducted for each outcome measure (first grade SAT-10, second grade SAT-10, and second grade September ORF). One analyses entered first grade April ORF level as the first predictor and first grade ORF slope as the second predictor. Another analyses reversed the order entry of these two variables. Regarding the slope estimates used, estimates from the linear and quadratic model were analyzed separately. A total of 12 separate regression analyses were conducted, which were broken down as three outcome measures, two predictor order entries, and two types of growth estimates. Overall, the results from all of these regression analyses indicated the ORF level in first grade was a strong predictor of reading outcomes. ORF growth, however, provided little or no unique contribution to the prediction.

Baker and colleagues (2008) examined the predictive validity of ORF growth, above initial ORF level, for performance on high-stakes reading assessments. The participants included four cohorts of first through third graders from Reading First schools with a total of over 9,600 students. DIBELS ORF was given during fall, winter, and spring. During each screening, the median of three passages was taken as the final score. The Stanford Achievement Test (SAT-10) was given in the spring of first and second grade. The Oregon Statewide Reading Assessment (OSRA) was given in the spring of third grade. Correlational analyses yielded the following: .72 to .82 (between first grade winter ORF to first grade spring ORF and first grade SAT-10), .63 to .80 (between first grade winter ORF to second grade spring ORF and second grade SAT-10), and .58 to .68 (between second grade fall ORF to third grade spring ORF and third grade OSRA). Generally, the higher correlations corresponded to ORF administrations that
were closer in time proximity to the administration of the SAT-10 or OSRA. Following correlational analyses, prediction models were constructed to examine the contribution of previous year’s high-stakes test score, ORF level, and ORF slope in predicting the current year’s high-stakes test score. One prediction model used first grade data to predict second grade SAT-10. Another prediction model used second grade data to predict third grade OSRA. For both models, it was found that all predictors (prior year’s high-stakes test score, ORF level, and ORF slope) significantly predicted performance on the current year’s high-stakes test. Regarding the specific contribution of ORF slope, it was found that it contributed 10% of the variance (out of a total of 76%) for predicting second grade SAT-10 and 3% of the variance (out of a total of 59%) in predicting third grade OSRA. Although ORF growth did contribute significantly to the prediction models, the practical significance of using ORF growth to predict third grade outcomes is questionable in light of the finding that it only contributed an additional 3% of the variance.

Stage and Jacobsen (2001) examined the validity of ORF to predict performance on the Washington Assessment of Student Learning (WASL). The participants included 173 fourth graders from one school. Of the participants, 90% were Caucasian, 15% were from low socioeconomic backgrounds, and 11 had special needs. ORF probes obtained from a district basal were given in September, January, and May. The WASL was given in May. Correlational analyses showed that ORF, given at all time points, was significantly related to WASL performance with correlations ranging from .43 to .51. Growth curve analysis was conducted on the WASL to predict ORF slope. Multiple regression analyses showed that ORF level was a significant predictor of WASL
performance. However, with ORF level in the model, ORF growth did not contribute significant variance to the prediction. A series of classification analyses was conducted on the September ORF cut-score of 100 words correct per minute. This cut-score yielded .66 sensitivity, .76 specificity, .41 positive predictive power (over a base rate of .20), .90 negative predictive power (over a base rate of .80), and .74 overall classification accuracy. Cohen’s kappa was found to be at .34, which indicates that the diagnostic efficiency of ORF was 34% above chance. In their discussion, the authors concluded that this study supports the use of ORF level for screening students at-risk for poor reading outcomes. The authors speculated that one reason that ORF growth did not contribute unique variance might be that ORF growth tends to level off in the upper elementary grades. ORF growth may be a better predictor for younger students.

Keller-Margulis and colleagues (2008) examined the validity of using ORF level and growth to predict performance on high-stakes reading assessments one or two years later. The participants included 1,461 elementary school students who were followed longitudinally. Of the participants, 8% were English learners, 33% were from low socioeconomic backgrounds, 58% were Caucasian, and 31% were Hispanic. ORF from AIMSweb were given in the fall, winter, and spring of first through fifth grade. The Pennsylvania System of School Assessment (PSSA) was given in the spring of third and fifth grade. The TerraNova Achievement Test was given in fourth grade. In analyzing the data, ORF level and slope values were correlated with PSSA or TerraNova outcomes one or two years later. ORF level was found to be moderately correlated with the PSSA and TerraNova. ORF slope in first grade was moderately correlated to PSSA outcomes in
third grade. However, ORF slope in fourth grade was not related to PSSA outcomes in fifth grade. Following correlational analyses, ROC analyses were used to identify cut scores for ORF level and growth in predicting reading outcomes. The following results were found for using ORF level to predict reading outcomes one or two years later in third, fourth, and fifth grade: .71 to .79 sensitivity, .78 to .91 specificity, .55 to .81 positive predictive power, .87 to .90 negative predictive power, and .76 to .81 overall classification accuracy. The following results were found for using ORF growth to predict reading outcomes two years later (first grade predicting third grade; second grade predicting fourth grade): .61 and .74 sensitivity, .61 and .81 specificity, .46 and .60 positive predictive power, .74 and .89 negative predictive power, and .61 and .79 overall classification accuracy. For these results using growth, the higher values correspond to predicting outcomes in third grade, while the lower values correspond to predicting outcomes in fourth grade. In their discussion and consistent with Baker and colleagues (2008), the authors conclude that the relationship between ORF growth and later reading outcomes is stronger in the earlier grades and decreases in the later grades.

Several themes can be drawn from studies described above examining the role of ORF growth in the prediction of later reading outcomes (Baker et al., 2008; Keller-Margulis et al., 2008; Schatschneider et al., 2008; Stage & Jacobsen, 2001). In all four studies, ORF level was found to be a significant predictor of performance on high-stakes state assessments or public norm-referenced tests. This result was found across various grades (first through fifth). However, mixed results were found for the predictive validity of ORF growth for later reading outcomes. Two of the studies suggest that ORF growth
has more predictive validity in the lower primary grades (Baker et al., 2008; Keller-Margulis et al., 2008). Schatschneider and colleagues (2008), however, found that ORF growth contributed little to no unique variance in predicting reading outcomes even in first grade. Stage and Jacobsen (2001) found that ORF growth did not contribute significant variance to predicting outcomes among fourth graders. Keller-Margulis and colleagues (2008) found that ORF growth in fourth grade was not related to outcomes in fifth grade. Overall, research supports the predictive validity of ORF level but yields mixed results concerning the predictive validity of ORF growth (Baker et al., 2008; Keller-Margulis et al., 2008; Schatschneider et al., 2008; Stage & Jacobsen, 2001). A review of general screening issues by Jenkins and colleagues (2007) also found mixed results regarding the utility of using growth data in screening.

**Using additional measures with ORF.** To improve screening accuracy, research has examined value of using various screening measures in addition to ORF to predict later reading outcomes (Wiley & Deno, 2005; Shapiro, Solari, & Petscher, 2008; Johnson et al., 2009; Johnson et al., 2010). These studies have typically examined the predictive contribution of reading comprehension, vocabulary, and other screening measures above and beyond the contribution of ORF. One advantage to using various screening measures is that it may increase the reliability of screening because multiple measures are being used. It may also help give more insight to the needs of students. However, a disadvantage of this approach is that the resources, both time and money, required to administer these additional screeners may challenge the practicality of their use (Jenkins et al., 2007). Another disadvantage is that, even if resources were not an issue,
interpretation of results becomes much more complex when using multiple measures in screening (Jenkins et al., 2007).

Johnson and colleagues (2009) sought to maximize screening accuracy by considering the use of a variety of screening measures. The participants of this study were 12,055 students in the primary grades. Of the participants, 42% were Caucasian, 30% were Black, 22% were Hispanic, 10% were English language learners, 21% were students with special needs, and 64% were students from low socioeconomic backgrounds. Students were assessed with DIBELS Initial Sound Fluency (ISF), Phoneme Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), ORF, and the Peabody Picture Vocabulary Test (PPVT) in kindergarten and/or first grade. The SAT-10 reading comprehension subtest was given at the end of first grade. The Florida Comprehensive Assessment Test – Sunshine State Standards (FCAT-SSS) was given in third grade as a measure of reading performance. Among various end-of-first-grade measures, the SAT-10 was selected as the first grade outcome measure because it yielded the highest correlation with third grade FCAT-SSS. Logistic regression was used to examine the predictive contribution of ISF, PSF, NWF, ORF, and PPVT. Then, the diagnostic utility of the measures were assessed through ROC analyses with sensitivity set at 90%. Among these measures, ORF taken in the fall of first grade was found to have the highest classification accuracy in predicting end-of-first-grade performance on the SAT-10. Despite being found as the most accurate measure, ORF yielded only 59% specificity when sensitivity was set at 90%. It was speculated that the low specificity may be due to the floor effects of ORF in the fall of first grade. In an attempt to reduce the
number of false positives, the researchers examined the value of considering first grade NWF and kindergarten PPVT scores in addition to ORF. However, it was found that adding either of these two measures improved sensitivity by less than two percent. Overall, this study found that adding additional screeners to ORF in first grade did not improve predictive accuracy in a practically significant way.

Johnson and colleagues (2010) examined the effect of combining various screening measures in order to maximize accuracy in predicting third grade reading performance. The participants were over 12,000 students including 41% Caucasians, 29% Blacks, 22% Hispanics, 17% English language learners, and 14% students with special needs. The outcome measure was the Florida Comprehensive Assessment Test – Sunshine State Standards (FCAT-SSS) given in February of third grade. The predictors under examination included following: DIBELS ORF given spring of second grade and fall of third grade, the PPVT given in February of second grade, and the SAT-10 given in February of second grade. Simultaneous logistic regression and ROC analyses were used to examine the predictors. It was found that the use of a single predictor resulted in fairly low classification accuracy. With sensitivity set at 90%, spring of second grade ORF yielded 43% specificity and 58% overall classification accuracy. Fall of third grade ORF yielded 45% specificity and 59% overall classification accuracy. Although using ORF as a single predictor yielded low classification accuracy, using a combination of predictors increased specificity to 68% and overall classification accuracy to 75%. This combination of predictors included spring of second grade ORF, second grade PPVT, second grade SAT-10, fall of third grade ORF, special education status, and English language learner
status. Overall, this study found that DIBELS ORF had low classification accuracy for predicting third grade reading outcomes. The authors recommend using a combination of screening measures to maximize prediction.

Shapiro and colleagues (2008) examined the value of adding a reading comprehension measure to ORF in predicting performance on a high-stakes state assessment in the area of reading. The participants included 1,000 third through fifth graders. Of the participants, 48% were Caucasian, 34% were Black, 12% were Hispanic, and 40% were from low socioeconomic backgrounds. DIBELS ORF and the 4Sight Benchmark Assessment were administered in the fall and winter. The 4Sight is a one-hour-long reading comprehension assessment that has a similar format as the statewide assessment. The Pennsylvania System of School Assessment (PSSA) was administered in the spring. Logistic regression and ROC analyses were used to in the data analysis. Regarding the use of DIBELS ORF, risk-status appears to moderate predictive validity. The strongest predictions were made for students scoring at or above the ORF benchmark. However, less accurate predictions were made for students in the ORF “at-risk” and “some-risk” categories with the least accurate predictions belonging to the “some-risk” group. Overall, it was found that ORF showed adequate predictive validity, but adding the 4Sight measure enhanced prediction for students in all “risk” categories. In the discussion, the authors suggest that adding a comprehension measure, such as 4Sight, would probably be the most useful for predicting the reading outcomes for students who score in the “some-risk” category for ORF.
Wiley and Deno (2005) compared the value of adding a comprehension measure to ORF in predicting reading performance on a high-stakes state assessment for native English speakers and English language learners. The participants included 36 third graders and 33 fifth graders. About 40% of the participants were ELLs (80% Hmong, 13% Somali, and 7% Spanish). ORF passages from the ‘Standard Reading Passages’ were administered every two weeks throughout the year. Maze, a one-minute cloze measure of reading comprehension, was administered in the fall, winter, and spring. The reading portion of the Minnesota Comprehensive Assessment (MCA) administered in the spring was used as the outcome measure. In this study, it was unclear which administration (fall, winter, or spring) of ORF and Maze was used in the analysis. Correlational analyses yielded values of .57 to .71 between ORF and MCA, and values of .52 to .73 between Maze and MCA. Multiple regression analyses indicated that when ORF was entered first and Maze entered second into the equation, Maze accounted for significant variance beyond ORF only for non-ELLs. For third grade non-ELLs, ORF contributed to 51% and Maze contributed an additional 11% of the variance in predicting MCA. For fifth grade non-ELLs, ORF contributed to 32% and Maze contributed an additional 22% of the variance. A second regression analyses was conducted by first entering Maze and then entering ORF into the equation. From this analysis, it was found that ORF did not account for significant variance in MCA for any of the students. Overall, this study suggests that both ORF and Maze are predictive of MCA performance in third and fifth grade, with ORF being a slightly better predictor. The authors conclude
that Maze becomes a more useful predictor of MCA as students become older and more proficient in English.

Taken as a whole, mixed results were found in studies described above examining the unique contribution of other screening measures above and beyond ORF in the prediction of later reading outcomes (Wiley & Deno, 2005; Shapiro et al., 2008; Johnson et al., 2009; Johnson et al., 2010). The results from Johnson and colleagues (2010) and Shapiro and colleagues (2008) support the use of other screeners in addition to ORF. Johnson and colleagues (2010) recommended using a combination of screening measures to maximize prediction of third grade reading outcomes. In addition to ORF, this combination of predictors should include a measure of receptive vocabulary, a reading achievement test, special education status, and English language learner status. Shapiro and colleagues (2008) suggested adding a reading comprehension measure (e.g.: 4Sight) to enhance prediction for third through fifth graders. Despite these studies supporting the use of additional screeners, Johnson and colleagues (2009) found that adding additional screeners (e.g. NWF and PPVT) to ORF in first grade did not improve predictive accuracy in a practically significant way. Wiley and Deno (2005) found that adding a comprehension measure to ORF only accounted for significant variance for native English speakers but not for ELLs in third and fifth grade.

**Other methods.** As described above, researchers have sought to improve the accuracy of screening by considering growth data and using multiple screeners. In addition to these methods, several others have also been examined in the literature including dynamic assessment (Jenkins et al., 2007) and the incorporation of teacher
judgment as a part of screening (Flynn & Rahbar, 1998; Speece et al., 2010). Dynamic assessment is an assessment method that involves teaching during the assessment. One advantage of dynamic assessment is that it yields information about student learning as part of the assessment. However, a disadvantage is that increased time is required for screening (Jenkins et al., 2007). A review of the research on screening has yielded mixed results regarding the effectiveness of dynamic assessment (Jenkins et al., 2007).

Regarding the incorporation of teacher judgment as a part of screening, limited research has been done in this area. The following is a discussion of the literature in the general area of teacher judgment. Afterwards, the discussion will turn to the added value of combining teacher judgment with other screeners.

Teacher Judgment

In a review, Hoge and Coladarci (1989) synthesized the literature examining teacher judgments of academic achievement. This review contained 16 studies that compared teacher judgments with standardized assessments of student achievement. Nine of these studies examined “indirect” judgments where teacher judgments of overall achievement were compared to student test scores. The other seven studies examined “direct judgments” where teachers were asked to estimate specific student test scores. The teacher judgments measures in these 16 studies varied in type and degree of specificity. Some measures required teachers to make peer comparisons while others just asked teachers about the target child. The following are examples of some of the measures: estimating grade equivalents, estimating specific item answers, giving general ratings, estimating total scores, and rank ordering on specific skills. Correlations and
regression coefficients were used in the majority of the studies to measure accuracy of judgments. For indirect teacher judgments, correlations ranged from .28 to .86 with a median of .62. For direct teacher judgments, they ranged from .48 to .92 with a median of .69. There was no significant difference between judgments that did and did not use peer comparisons. Most of the studies in this review did not examine teachers as a source of variability. Among the four studies that did, some evidence of teacher variability in judgments was found. Four studies compared judgments of students from different achievement levels. Three of these four studies found that teachers made less accurate judgments for lower achieving students (vs. high achievement students). From this review, the authors concluded that teacher judgments of student performance were fairly valid. One final point to note in this review is that the teacher judgments were for concurrent student performance rather than for predicting future performance.

In a later review, Coladarci (1992) examined studies of teacher judgments specific to student reading performance on standardized assessments. Three general types of teacher judgments studies were found in this review. One group of studies asked teachers to look through specific reading assessments and identify the items they predict would be answered correctly. This was done for each student. Teachers’ degree of judgment accuracy was determined by calculating the proportion of test items that were judged correctly. A second group of studies asked teachers to predict each student’s score on specific reading assessments. These scores included grade equivalents, raw scores, and percentiles. Teachers’ degree of judgment accuracy was measured using correlations between teacher judgments and students’ actual scores. A third group of studies asked
teachers to rate each student’s general reading skills independent of any specific reading assessment. This was typically done on a scale of one to five. Teachers’ degree of judgment accuracy was measured using correlations between ratings and students’ actual scores. Similar to the review by Hoge and Coladarci (1989) examining general academic performance, this study on reading performance found that teachers made generally accurate judgments with an average correlation of .67 between various forms of teacher judgments and student performance. Again, some evidence of teacher variability in judgments was found, and judgments were less accurate for lower performing students. In addition to these reviews, individual studies have examined the relationship between teacher judgment and student achievement. The following is a discussion of some of these studies, organized first by academic area (general achievement vs. reading-specific) and then by chronological order.

**Judgments of general achievement.** In an early study, Coladarci (1986) examined the accuracy of teacher judgments of student performance on standardized assessments. The participants included eight third and fourth grader teachers. Six random students from each teacher’s class were selected as targets of judgment. For each student, teachers predicted student responses as either correct or incorrect on specific items on a standardized assessment of math and reading. The results indicated that teachers made accurate judgments for around 75% of the test items. Significant differences in judgment accuracy as a function of teacher variability was found. Also, teacher judgments were found to be more accurate with higher performing students than with lower performing students.
Gresham, MacMillan, and Bocian (1997) examined the accuracy of teacher judgments for identifying students who are at-risk for learning difficulties. The participants included 240 second through fourth graders and their teachers ($n = 60$). The students were divided into the following four groups: students with learning disabilities, students with low intelligence, students with low achievement, and students with typical achievement. Teachers completed the Academic Competence Scale (from the Social Skills Rating System) for each student, which rated students in the areas of cognition, motivation, math, reading, parent support, and classroom behavior relative to other peers ($1 = $lowest $10\%$, $5 = $highest $10\%$). It was found that teachers were accurate in differentiating typically achieving students from all other students. However, teachers had difficulty differentiating among students with learning disabilities, students with low intelligence, and students with low achievement.

Demaray and Elliot (1998) examined teacher judgment accuracy with a sample of 12 teachers and 47 students in first through fourth grade. Student achievement was measured using the Kaufman Test of Educational Achievement (K-TEA), which assesses in the areas of math, reading, and spelling. Teacher judgment was measured using the Academic Competence Scale (from the Social Skills Rating System) and a questionnaire asking teachers to predict student responses on each item of the K-TEA as correct or incorrect. A .70 correlation was found between K-TEA performance and teacher judgments on the Academic Competence Scale. A slightly higher correlation of .84 was found between K-TEA performance and the teacher questionnaire. Also, when using the teacher questionnaire, more accurate judgments were made for above-average students.
than for below-average students. However, this result was not found when using the Academic Competence Scale. Regarding teacher variables, teaching experience did not influence judgment accuracy.

Judgments of reading achievement. Bates and Nettelbeck (2001) compared teacher judgments with student performance in reading comprehension and accuracy. The participants included 108 students age six through eight and their teachers ($n = 29$). The Neale Analysis of Reading Ability – Revised (NARA-R) was used to assess students’ reading skills. Teachers were asked to assign a percentile rank to each student in the areas of reading comprehension and accuracy. Teacher judgments were found to correlate moderately with both student reading comprehension ($r = .62$) and accuracy ($r = .77$). However, teacher judgments were found to tend towards the middle (e.g.: over-estimating skills of lower readers; under-estimating skills of higher readers). Also, teachers were more accurate at making relative rankings of students rather than making absolute judgments of student performance. Regarding variation among teachers, number of years teaching did not affect judgment accuracy.

Feinberg and Shapiro (2003) examined the accuracy of teacher judgments of students’ reading fluency. The participants included 30 third through fifth grade teachers and 30 students (one randomly selected from each teacher’s class). Students were administered three ORF passages and the median was taken as the final score. Teachers were given the ORF passages and asked to predict their student’s score on each passage. The median of three predictions was taken as the final decision. Teachers were also asked to rate students using the reading/language arts subscale of the Academic Competence
Evaluation Scales (ACES), which contained five items with possible ratings of “one” through “five.” Correlational analyses found a significant relationship between student performance and teacher estimates of ORF scores ($r = .70$) as well as teacher ratings on the ACES ($r = .62$). The two correlations were not significantly different. Judgment accuracy was not found to vary by number of years teaching.

Similar to the previous study, Madelaine and Wheldall (2005) examined teacher judgments of students’ reading skill as measured by ORF. The participants included 33 third through fifth grade teachers and 396 students (12 randomly selected from each class). Teachers were asked to rank order their students based on general reading skills (1 = highest student, 12 = lowest student). The researchers rank ordered students based on their ORF scores. Results yielded an overall correlation of .73 (range .32 to .99) between teacher rankings and ORF rankings. However, disagreements were found between these two sources. For example, 30% of teachers ranked students in the top 25% when ORF ranked them in the bottom 50%. Also, 24% of teachers ranked students in the bottom 25% when ORF ranked them in the top 50%. Despite this, it is promising that over half of the teachers agreed with ORF in terms of identifying the lowest reader in the class.

Another study by Begeny, Eckert, Montarello, and Storie (2008) also examined teacher judgments of students’ reading skill as measured by ORF. The participants included 87 first through third graders and their teachers ($n = 10$). Teacher judgment was measured using a variety of methods including the Teacher Rating Scale of Reading Performance (TRSRP), a teacher interview, and class ranking charts. The TRSRP asked teachers to rate students reading skills on a Likert-type scale. The teacher interview sheet
asked teachers to estimate students’ number of words read correctly/incorrectly per minute and students’ instructional levels. The class ranking chart asked teachers to rate each student’s reading accuracy and fluency as compared to those of classroom peers. Correlations between students’ ORF performance and all judgment methods ranged from .53 to .79. Although the accuracy of teacher judgments was found to vary depending on methods of data analyses, a general finding was that teachers were less accurate judging the performance of students with low to average ORF performance (compared to students with high ORF performance). Also, teachers gave more accurate judgments in the form of ranking students relative to their peers than they did with direct measures of judgment (e.g.: estimating number of words read correctly). Lastly, teachers did not vary in accuracy among themselves.

Feinberg and Shapiro (2009) compared teacher judgments of students reading skills among students’ with varying levels of performance. The participants included 74 second to fifth grade teachers. One average-performing student and one low-performing student were selected as participants from each teacher’s class. Each student was administered ORF and the Woodcock-Johnson III Tests of Achievement (Letter-word Identification and Reading Comprehension subtests). Teachers were asked to estimate their students’ ORF scores and rate their students relative to peers using the Academic Competence Evaluation Scale (ACES). Correlations between .47 and .60 were found between student performance and teacher judgments. Conflicting results were found in regards to judgment accuracy as a function of student performance. When asked to predict ORF scores, teachers were less accurate at making predictions for lower-
performing students. However, when asked to rate students relative to peers using the ACES, teachers displayed the same levels of accuracy with all students.

**Moderators of teacher judgment accuracy.** The reviews and studies discussed previously varied in terms of the type of teacher judgment measures employed. In the review by Hoge and Coladarci (1989), about half of the studies used “direct” measures where teachers were asked to estimate specific test scores or responses to specific test items. The other half used “indirect” measures of judgments where teacher judgments of overall achievement or judgments using peer comparison were compared to student test scores. For the individual studies, only Coladarci (1986) measured teacher judgments directly in the form of estimating responses to specific test items. Two of the individual studies used indirect measures with peer comparison (Gresham et al., 1997; Madelaine & Wheldall, 2005). For the rest of the individual studies, both direct and indirect measures were used (Bates & Nettelbeck, 2001; Begeny et al., 2008; Demaray & Elliot, 1998; Feinberg & Shapiro, 2003; Feinberg & Shapiro, 2009). Based on this discussion, both types of measures seem to be used with relatively equal frequency in the teacher judgment literature. Regarding accuracy of teacher judgments as a function of measurement type, some studies suggest that teachers are more accurate at making relative judgments rather than making absolute judgments of student performance (Bates & Nettelbeck, 2001; Begeny et al., 2008).

In addressing the role of teacher variability on accuracy of teacher judgments, both reviews by Hoge and Coladarci (1989) and Coladarci (1992) found judgment accuracy varied by teacher. Coladarci (1986) found similar results. However, these
studies did not explore the source of variability that contributed to differences among teachers. Among other studies that did explore this issue, it was found that number of years in teaching did not influence judgment accuracy (Bates & Nettelbeck, 2001; Demaray and Elliot, 1998; Feinberg & Shapiro, 2003). One study by Begeny and colleagues (2008) found that teachers did not vary in accuracy of judgments.

Regarding the role of student performance on teacher judgment, the reviews by Hoge and Coladarci (1989) and Coladarci (1992) both found that teachers generally made less accurate judgments for lower-performing students than they did for higher-performing students. This finding was also present in some individual studies that examined the issue (Begeny et al., 2008; Coladarci, 1986). Other studies found more mixed results. For example, both Demaray and Elliot (1998) and Feinberg and Shapiro (2009) found teachers to make more accurate judgments with higher-performing students only when they were providing “direct” judgments (e.g.: estimating a score on a certain test). However, this effect was not present when teachers were asked to make relative comparisons using a rating scale. In another study that examined this issue, no effect was found (Bates & Nettelbeck, 2001). For studies for which student performance made a difference, teachers always tended to be less accurate with lower-performing students and more accurate with higher-performing students.

**Incorporating Teacher Judgment to Screen for At-risk Readers**

A comparatively limited number of studies in the literature have sought to improve screening accuracy by combining teacher judgment with other screening measures. After conducting their study on teacher judgments of students reading skills,
Feinberg and Shapiro (2009) suggested in their discussion that the issue of combining curriculum-based measurement with teacher judgment to improve screening needs to be explored. A study by Flynn and Rahbar (1998) explored the effect of combining a teacher rating scale with a literacy screening battery. The participants were 210 kindergarteners, most of whom were Caucasian. All students were screened using a researcher-developed literacy screening battery that evaluated 10 early literacy skills. Each student’s teacher also completed a research-developed 10-item rating scale with each item corresponding to a domain in the screening battery. Students were identified as “at-risk” if they scored below the 35\textsuperscript{th} percentile in either measure. The outcome, student reading performance, was measured using reading scores from numerous standardized assessments (e.g.: ITBS, MAT, SAT). Poor later outcomes were defined as scoring at or below the 40\textsuperscript{th} percentile in first, second, or third grade. The results indicated that the literacy screening battery yielded 80\% sensitivity and 72\% specificity. The teacher rating scale yielded 64\% sensitivity and 86\% specificity. When the two measures were combined, sensitivity increased to 88\%. However, specificity decreased to 57\%. In their discussion, the authors suggested that teacher ratings should be combined with screening assessments in order to maximize sensitivity.

In another study, Speece and colleagues (2010) evaluated various measures of student achievement and teacher ratings regarding their utility in identifying students who are at risk for poor reading outcomes. The participants included 230 fourth graders from 20 classrooms who were 74\% Caucasian, 18\% Black, and 4\% with special needs. The students were given a number of group and individually administered reading
assessments. These included measures of oral language, auditory memory, phonological processing, word decoding, word recognition, reading comprehension, and spelling. The teachers were asked to competing three rating scales for each child. These included the Academic Competence Scale (from the SSRS), the ADHD Rating Scale, and the Teacher Reading Rating Form (give a rating of 1 – 5 for each child’s reading level). Students were classified as not at-risk or at-risk based on a three-factor model that included decoding, fluency, and comprehension. At-risk was defined as having a score below the 15th percentile in any factor. All-subsets regression was used to identify measures that most strongly and efficiently predicted risk status.

The predictors were evaluated in the following order: 1) group administered tests and teacher ratings, 2) individually administered tests, and 3) growth scores. For the first group of predictors, ones found to be most optimal included a measure of reading comprehension (Gates-MacGinitie Reading Test; GMRT), silent word reading fluency (Test of Silent Word Reading Fluency; TOSWRF), and teacher ratings (Teacher Rating of Reading Problems). These measures accounted for 46% of the variance. Two individually administered tests were found to contribute statistically, but not practically, significant additional variance. None of the growth scores contributed significant variance. After the identification of a set of predictors, classification accuracy was assessed using ROC analysis. With sensitivity set at .80, specificity was found to also be .80. This curve yielded an AUC of .90. However, with sensitivity set at .90, specificity decreased to .63. From this study, the authors concluded that multiple measures should be
considered when screening for students at-risk for reading difficulties in the later elementary grades.

**Gaps in the Screening and Teacher Judgment Literature**

In considering the above studies, it can be suggested that current screening tools possess limited predictive validity and accuracy in identifying students at-risk for reading difficulties. The predictive accuracy of screening tools do not meet the guidelines of 90% sensitivity as recommended in the recent screening literature (Johnson et al., 2009; 2010). Research has examined various methods in an effort to increase the predictive validity and accuracy of the screening process. These methods have included the use of growth data, multiple screening measures, and dynamic assessment. However, research has yielded mixed results regarding the adequacy of these methods (Baker et al., 2008; Jenkins et al., 2007; Johnson et al., 2009, 2010; Keller-Margulis et al., 2008; Schatschneider et al., 2008; Shapiro et al., 2008; Stage & Jacobsen, 2001; Wiley & Deno, 2005). Regarding the incorporation of teacher judgment as a part of screening, limited research has been done in this area. Therefore, there is a clear need to conduct more research in this area in order to inform improvements in the predictive validity and accuracy of screening tools.

Although there are some studies examining the relationship between teacher judgments and reading performance (e.g.: Bates & Nettelbeck, 2001; Begeny et al., 2008; Feinberg & Shapiro, 2003; Feinberg and Shapiro, 2009; Madelaine & Wheldall, 2005), one limitation is that research has not explored the relationship between teacher judgment and reading performance as measured by high-stakes state accountability assessments. In
these existing studies, teacher judgment is typically compared with students’ concurrent reading performance rather than future performance. There are a few studies that examine the effect of adding teacher judgment to other screening measures (e.g.: Flynn & Rahbar, 1998; Speece et al., 2010). However, studies have not examined the effect of combining teacher judgment with curriculum-based measures (e.g. ORF) to screen for students at-risk for poor reading outcomes. The value of teacher judgment as a predictor of future reading performance needs further examination in order to explore its role as a screening tool for at-risk students.

Another limitation in the literature is that although there are numerous studies examining the effect of English language learner status on the predictive validity of ORF, these studies have yielded conflicting results (e.g.: Hosp et al., 2011; Klein and Jimerson, 2005; Johnson et al., 2009; Johnson et al., 2010; Muyskens et al., 2009; Roehrig et al., 2008; Wiley and Deno, 2005). One possible explanation for these mixed results is that studies have typically characterized ELLs as a homogeneous group or, at best, as separated by native language. Level of English language proficiency has not been considered as a source of variation. However, it is common knowledge that ELL students possess a wide range of English language skills (CDE, 2011). According to recommendations for test standards set forth by AERA, APA, and NCME (1999), language proficiency should be considered as a part of assessment and interpretation. Some researchers have suggested that ELLs may need different cut-scores than the ones used for native English speakers (Jenkins et al., 2007). Therefore, future research on the predictive validity of ORF needs to consider the influence of language proficiency.
A final limitation in the literature is that there is limited research examining student variables that influence the accuracy of teacher judgment. Current research has suggested student achievement as one of these variables. These studies have generally found teachers to be more accurate judges of higher performing students than of lower performing students (Begeny et al., 2008; Coladarci, 1986; Coladarci, 1992; Hoge & Coladarci; 1989). However, a few studies have also suggested that student achievement does not play a role in teacher judgment accuracy (Bates & Nettelbeck, 2001; Demaray and Elliot, 1998; Feinberg and Shapiro, 2009). In addition to student achievement, other student variables (e.g.: English language proficiency) influencing the accuracy of teacher judgment need to be explored.

Purpose of Study

To address the gaps in the literature as described above, this study focused on issues related to the predictive validity of curriculum-based measurement, use of teacher judgment with curriculum-based measures as a part of screening, and moderators of teacher judgment. The purpose of this study was to examine the effects of incorporating teacher judgment with curriculum-based measurement (e.g.: AIMSweb R-CBM and Maze) as a part of screening to identify students at-risk for poor reading performance on a high-stakes state accountability assessment. The predictive validity and accuracy of R-CBM, Maze, and teacher judgment were compared. In addition, the effect of student variables (e.g.: English language proficiency, level of reading achievement) and teachers on the contribution of teacher judgment was examined. The following research questions were explored with a sample of third grade students and a sample of fifth grade students:
1) What is the relationship between R-CBM, Maze, and teacher judgment with spring reading outcomes as measured by the CST-ELA-RC; and to what extent does English language proficiency moderate this relationship?

2a) After controlling for performance on prior year’s state test and performance on AIMSWEB screening measures (R-CBM and Maze), to what extent does teacher judgment contribute additional variance in predicting spring reading outcomes as measured by the CST-ELA-RC; and to what extent does English language proficiency affect this contribution?

2b) After controlling for performance on prior year’s state test and teacher judgment, to what extent does performance on AIMSWEB screening measures (R-CBM and Maze) contribute additional variance in predicting spring reading outcomes as measured by the CST-ELA-RC; and to what extent does English language proficiency affect this contribution?

3) What is the predictive accuracy of AIMSWEB cut-scores (R-CBM and Maze) and teacher judgment for spring performance on the CST-ELA-RC; and to what extent does English language proficiency affect predictive accuracy?

4) To what extent does level of student reading achievement affect the predictive contribution of teacher judgment in screening?

5) To what extent are teachers a source of variability in the accuracy of their judgments; and to what extent do teacher variables (e.g.: teaching experience and teaching qualifications) explain teacher variability?
Method

Participants

**Students.** The student participants of this study were selected using convenience sampling. The sample consisted of 1,035 third grade and 1,069 fifth grade students from 20 public elementary schools within one urban school district. The third grade sample included 50% \((n = 516)\) native English speakers, 16% \((n = 168)\) ELLs with early advanced to advanced English proficiency, and 34% \((n = 351)\) ELLs with beginning to intermediate English proficiency. The fifth grade sample included 57% \((n = 612)\) native English speakers, 23% \((n = 245)\) ELLs with early advanced to advanced English proficiency, and 20% \((n = 212)\) ELLs with beginning to intermediate English proficiency. All ELL students spoke Spanish as their primary language. Approximately 84% of students in the school district were from socio-economically disadvantaged backgrounds (e.g.: eligible for free-reduced lunch).

**Teachers.** The teacher participants of this study consisted of 52 third grade and 50 fifth grade general education teachers for the above student sample. This teacher sample represents 67% of the 152 teachers recruited to participate in the study. The recruitment process is described below in the procedures section. Regarding years of teaching experience, the third grade teachers ranged from 2 to 42 years \((M = 12.40, \text{median} = 11)\) and the fifth grade teachers ranged from 1 to 22 years \((M = 10.98, \text{median} = 11)\). Approximately 29% \((n = 15)\) of the third grade teachers and 10% \((n = 5)\) of the fifth grade teachers possessed additional certification to teach reading (e.g.: Reading/Language Arts Specialist Credential, Reading Certificate). Approximately 65%
(n = 34) of the third grade teachers and 58% (n = 29) of the fifth grade teachers possessed additional certification to teach ELL students (e.g.: Bilingual authorization, BCLAD, BCC, Bilingual Crosscultural Specialist Credential). In addition, approximately 85% of all teachers reported that, in considering all of their years teaching, greater than 50% of their students have been ELLs. The teacher ethnicity breakdown was as follows: 30% Black, 27% Hispanic, 19% White, 13% Asian, and 11% other.

**Measures**

**AIMSWEB Reading Curriculum Based Measurement (R-CBM).** AIMSWEB R-CBM is an individually administered standardized measure of reading fluency. In its administration, the examiner presents a grade-level reading passage to the student. The student’s task is to read the passage out loud. The student’s score is the number of words correctly read in one minute. Three R-CBM probes are administered and the median score is taken as the student’s final score. The following are the validity and reliability coefficients for third and fifth grade R-CBM passages as reported by the AIMSWEB Technical Manuel (Pearson, 2012): average alternate form reliability (.93 third grade; 94 fifth grade), average test-retest reliability (.93 third grade; .94 fifth grade), and criterion-related validity (.66 -.77 third grade; .67 -.69 fifth grade).

**AIMSWEB Maze.** AIMSWEB Maze is a three-minute standardized measure of reading comprehension that can be individually or group-administered. Students are presented with a passage with missing words that are replaced with a choice of three words. The students’ task is to circle the correct missing word. The student’s score on the Maze is the number of words correct. The following are the validity and reliability
coefficients for third and fifth grade R-CBM passages as reported by the AIMSWEB Technical Manuel (Pearson, 2012): average alternate form reliability (.70 third grade; .78 fifth grade) and criterion-related validity (.58 - .59 third grade; .54 - .58 fifth grade).

**Teacher Rating Scale of Reading (TRS-R).** The TRS-R (see Appendix A) is a researcher-developed teacher judgment measure that asked teachers to rate the reading skills of each student in their class on a five-point scale using the following criteria: far below basic (1), below basic (2), basic (3), proficient (4), and advanced (5). The development of the TRS-R was informed by the California state standards and performance descriptors for reading comprehension (California Department of Education, 2009). Specifically, the first page of the TRS-R presented teachers with reading standards and performance descriptors corresponding with the grade-level taught (California Department of Education, 2009). The second page of the TRS-R contained a list of student names corresponding to each teacher’s classroom. The numbers “1” through “5” were listed next to each student’s name to corresponded with their reading levels (e.g.: far below basic – advanced). Teachers were asked to base their ratings of each student on the following question: How does this student’s reading comprehension skills compare to that expected by grade-level standards?

**Teacher Demographics Survey (TDS).** The TDS (see Appendix A) is a researcher-developed survey used to collect teacher demographic information (e.g.: grade taught, ethnicity, number of years teaching, experience working with ELLs, and additional certification).
California English Language Development Test (CELDT). The CELDT (CDE, 2011) is a measure used to assess English language proficiency levels of ELLs. It is administered each year to determine progress in English language development. The CELDT assesses students in English in the areas of listening, speaking, reading, and writing. English proficiency levels range from 1 to 5 (1 = beginning, 2 = early intermediate, 3 = intermediate, 4 = early advanced, 5 = advanced). The content validity of the CELDT has been supported by educational experts who have aligned the test’s content to California’s English Language Development (ELD) standards. Convergent validity has been addressed by examining correlations among the CELDT scales ($r = .49$ to $.74$). Construct validity has been addressed through the test development process by minimizing construct underrepresentation and construct-irrelevant variance. Internal consistency reliability coefficients across all domains of the CELDT range from .71 to .91. Interscorer reliability coefficients for the writing component range from .65 to .98 (CDE, 2011).

California Standards Tests – English Language Arts – Reading Comprehension (CST-ELA-RC). For the current study, the CST-ELA-RC (CDE & ETS, 2011) was used as the outcome variable indicative of reading comprehension skills. The CST-ELA-RC is a subscale of the CST-ELA domain of the CSTs, which is the high stakes state accountability assessment used in California. The CST is administered each year to all public school students starting in second grade. This CST-ELA-RC subscale assesses performance in reading comprehension against California’s content standards. The content validity of the CST-ELA-RC was supported by educational experts during
the test development period. Convergent validity has been addressed by examining correlations between the CST-ELA and the California Achievement Test - Sixth Edition (CAT/6) Reading Language tests, which range from .75 to .80. Internal consistency reliability coefficients for the CST-ELA range from .93 to .95 (CDE & ETS, 2011).

Scores on the CST-ELA are categorized into the following categories: Far Below Basic, Below Basic, Basic, Proficient, and Advanced. A cut score of 350 or above is needed to score in the Proficient or Advanced categories. It is the California Board of Education’s goal for all students to score Proficient or Advanced (CDE, 2012). To aide in the interpretation of CST-ELA-RC scores, grade-specific cut-scores were derived that represent the “average percent correct for a representative sample of the state’s students who scored at the lowest score for proficient” on the CST-ELA (e.g.: scale score of 350). A CST-ELA-RC score of .73 for third grade and .70 for fifth grade correspond to minimum proficiency on the CST-ELA (CDE, 2012).

Procedure

District and school administrators provided access to the following student data: CELDT level, AIMSWEB R-CBM score, AIMSWEB Maze score, and CST-ELA-RC score. The data were collected by school staff as a part of routine educational assessment. CELDT levels were measured in September by school staff. AIMSWEB R-CBM and Maze scores were collected during the winter screening period (e.g.: third week of January) by classroom teachers, instructional aides, and graduate students consultants employed by the school district. Inter-rater reliability data for AIMSweb administrations collected from 19 teachers for this study and for a large study within the district yielded a
mean of 90.4% percent agreement. The following formula was used to calculate reliability: percent agreement = \[\text{agreements} / (\text{agreements} + \text{disagreements})\] x 100.

Reliability data were collected by trained graduate student consultants. CST-ELA-RC scores were collected in April by school staff. All assessments were administered using standardized procedures in a quiet setting at the students’ schools.

The TRS-R and TDS were distributed to all third and fifth grade teachers to complete during the winter screening period. Specifically, the teachers received a note regarding the study in their school mailboxes along with a copy of the informed consent, TRS-R, and TDS. The teachers completed these measures at their convenience. Of the 152 teachers recruited to participate, 67% completed and returned the TRS-R and TDS. As an incentive, all participating teachers were given a five-dollar gift card upon completion.

Prior to its distribution to the participating teachers, the TRS-R and TDS were piloted with two teachers. During the pilot process, an interview was conducted with each teacher after completion of the rating scale. The purpose of the pilot process was to ensure that teachers based their responses on the intended construct of the scale. The pilot process also helped ensure that the presentation of the measures was clear. Results from the pilot process indicated that both teachers perceived the TRS-R and TDS to be clear, organized, and self-explanatory. Both teachers based their responses on the intended construct of the scale. Therefore, no modifications were made to the TRS-R and TDS.
Results

Description of Data

The student participants consisted of 1,035 third grade and 1,069 fifth grade students. Only students with complete data were included in the study (e.g.: previous CST-ELA-RC, R-CBM, Maze, teacher judgment, and CST-ELA-RC). Table 1 contains the following student descriptive statistics organized by grade and ELL status: R-CBM, Maze, teacher judgment, and CST-ELA-RC score. The teacher participants consisted of 52 third grade and 50 fifth grade general education teachers of the student participants. For all research questions, analyses were conducted separately by grade.

Prior to the analyses, the following statistical assumptions were examined separately by grade and English proficiency status: linearity, normality, homogeneity of variance, independence, and lack of multicollinearity. These assumptions were examined for scores on previous CST-ELA-RC, R-CBM, Maze, teacher judgment, and CST-ELA-RC. Examination of linearity by use of scatter plots revealed linear relationships between scores on all measures and the CST-ELA-RC. Examination of normality by values of skewness and kurtosis revealed relatively adequate normality. Specifically, all variables yielded skewness and kurtosis values between +1 and –1 with the exception of Maze for ELLs (skewness range: 1.03 – 1.23; kurtosis range: 1.03 – 1.86). Examination of homogeneity of variance by use of residual plots revealed that this assumption was met. Examination of multicollinearity among previous CST-ELA-RC, R-CBM, Maze, and teacher judgment revealed no multicollinearity ($r < .70$) with the exception of two slight violations between R-CBM and teacher judgment ($r = .73$ for third grade native English
speakers; \( r = .72 \) for third grade as a whole). The assumption of independence was likely violated due to nested data (e.g.: students within classrooms). This violation was addressed in the last research question, which used multi-level modeling.

**Research Question 1**

To address the first research question, Pearson’s correlations were used to examine the relationship between R-CBM, Maze, and teacher judgment with spring reading outcomes as measured by spring CST-ELA-RC. The extent to which English language proficiency affects this relationship was examined by conducting separate analyses for each group. For these analyses, participants were categorized as native English speakers, ELLs at early advanced to advanced English proficiency, or ELLs at beginning to intermediate English proficiency. The results of the correlational analyses are reported in Table 2. For both grades, correlations between all measures and CST-ELA-RC for ranged from \( r = .22 \) to \( r = .61 \). The correlations ranged from \( r = .32 \) to \( r = .61 \) for R-CBM and CST-ELA-RC, \( r = .22 \) to \( r = .38 \) for Maze and CST-ELA-RC, and \( r = .27 \) to \( r = .60 \) for teacher judgment and CST-ELA-RC. Regarding the type of measure, Maze exhibited the lowest correlation with spring CST-ELA-RC across all grades and language proficiency levels. Specifically, the relationship between Maze and CST-ELA-RC was non-significant for all ELLs except third grade ELLs at beginning to intermediate English proficiency.

After examining the correlations between each measure and CST-ELA-RC, the correlations between each language proficiency group were compared for significant differences. The extent to which there was a significant difference between the
correlations of each group were examined by comparing the correlations using the following method as described by Cohen, Cohen, West, and Aiken (2003). All correlations being compared were transformed into $z$-scores using the following equation: 

$$Z_r = \frac{\ln(1+r) - \ln(1-r)}{2}.$$ 

The null hypothesis (no significant difference between the correlations) for this comparison was represented as $H_0: Z_{\rho_1} - Z_{\rho_2} = 0$. The alternative hypothesis (significant difference between the correlations) was represented as $H_a: Z_{\rho_1} - Z_{\rho_2} \neq 0$. The estimate or difference between the two correlations ($z_{r1} - z_{r2}$) being compared was then be divided by the standard error of the estimate $\sqrt{1/(n_1-3) + 1/(n_2-3)}$ to yield a standardized $z$-score. The $p$-value of this corresponding $z$-score was used to determine if the correlations exhibited a significant difference. The results of the comparisons are reported in Table 3.

In examining the correlations between R-CBM and CST-ELA-RC for third grade students, it was found that the correlation for ELLs at early advanced to advanced English proficiency ($r = .32$) was significantly lower than those for native English speakers and ELLs at beginning to intermediate English proficiency ($r = .58$ and $r = .49$, respectively). However, when comparing native English speakers and ELLs at beginning to intermediate English proficiency, there was no significant difference present.

Regarding the correlations between Maze and CST-ELA-RC for third grade students, it was found that the correlation for native English speakers ($r = .38$) was significantly higher than that for ELLs at beginning to intermediate English proficiency ($r = .25$), which was in turn higher than that for ELLs at early advanced to advanced English proficiency ($r = .08$, $p > .05$). For the correlations between CST-ELA-RC and teacher
judgment for third grade students, the correlation for native English speakers \((r = .58)\) was significantly higher than those for each of the ELL language proficiency groups \((r = .36\) and \(r = .43\)). However, the correlations for the ELL groups were not significantly different from each other. In examining the correlations between CST-ELA-RC and all measures for fifth grade students, it was found that the correlation for native English speakers \((r = .56)\) was significantly higher than those for each of the ELL language proficiency groups \((r = .31\) and \(r = .27\)). The correlations for the ELL groups were not significantly different from each other.

**Research Question 2**

**Research Question 2a.** The predictive validity of teacher judgment after controlling for performance on prior year’s CST-ELA-RC and AIMSWEB was examined using hierarchical linear regression. The following predictors were entered separately in the following order: 1) previous year’s CST-ELA-RC score, 2) R-CBM score, 3) Maze score, and 4) teacher judgment score. The order of variable entry was informed by logic and prior research. Specifically, previous year’s CST-ELA-RC score was entered first as the control primary control variable. Teacher judgment was entered last in order to examine its unique predictive contribution. R-CBM score was entered before Maze due the comparatively larger amount of research support for its predictive validity. The outcome variable was the current year’s spring CST-ELA-RC score. The effect of English language proficiency was examined by conducting separate analyses for each group. For these analyses, participants were categorized as native English speakers, ELLs at early advanced to advanced English proficiency, or ELLs at beginning to
intermediate English proficiency. The statistical significance of each predictor was examined following its entry into the regression model. Non-significant predictors were not retained. The proportion of variance explained was examined following the entry of each significant predictor. See Table 4 for a summary of the proportion of variance explained by each predictor.

The following is a summary of notable findings regarding the predictive contribution of teacher judgment. After controlling for performance on prior year’s CST-ELA-RC and performance on AIMSWEB screening measures, teacher judgment was found to contribute additional variance in predicting spring CST-ELA-RC scores. The only group for which teacher judgment did not contribute additional variance was for fifth grade ELLs at beginning to intermediate English proficiency. The amount of additional variance contributed by teacher judgment ranged from 2% to 6%. As depicted in Table 4, the amount of additional variance contributed varied by English language proficiency. Among third grade students, teacher judgment contributed an additional 6% variance for ELLs at early advanced to advanced English proficiency, 3% variance for native English speakers, and 2% variance for ELLs at beginning to intermediate English proficiency. Among fifth grade students, teacher judgment contributed an additional 4% variance for native English speakers and 2% variance for ELLs at early advanced to advanced English proficiency.

Aside from the findings regarding teacher judgment, several other findings were noted. Across both grades, the combination of all predictors contributed a much greater proportion of variance for native English speakers (total $R^2 = .46$ and .49) than for ELLs.
(total $R^2$ ranging from .12 to .32). This contrast between native English speakers and ELLs was primarily due to the predictive contribution of previous year’s CST-ELA-RC performance. For native English speakers, the contribution of previous CST-ELA-RC was $R^2 = .32$ for third grade and $R^2 = .37$ for fifth grade. For ELLs, these values ranged from $R^2 = .03$ to $R^2 = .22$. A final notable finding was that Maze was found to be a significant predictor after controlling for previous CST-ELA-RC and R-CBM only for third grade ELLs at early advanced to advanced English proficiency ($R^2 = .03$).

**Research Question 2b.** Part b of the second research question examined the predictive validity of AIMSWEB performance after controlling for performance on prior year’s performance on the CST-ELA-RC and teacher judgment. Similar to the previous research question, hierarchical linear regression was used for these analyses. The following predictors were entered separately in the following order: 1) previous year’s CST-ELA-RC score, 2) teacher judgment score, 3) R-CBM score, and 4) Maze score. The outcome variable was the current year’s spring CST-ELA-RC score. The effect of English language proficiency was examined by conducting separate analyses for each group (e.g.: native English speakers, ELLs at early advanced to advanced English proficiency, and ELLs at beginning to intermediate English proficiency). The statistical significance of each predictor was examined following its entry into the regression model. Non-significant predictors were not retained. The proportion of variance explained was examined following the entry of each significant predictor. See Table 5 for a summary of the proportion of variance explained by each predictor.
The following is a summary of notable findings regarding the predictive contribution of R-CBM and Maze. After controlling for performance on prior year’s CST-ELA-RC and teacher judgment, R-CBM was found to contribute additional variance in predicting spring CST-ELA-RC scores. The only group for which R-CBM did not contribute additional variance was for third grade ELLs at early advanced to advanced English proficiency. The amount of additional variance contributed by R-CBM ranged from 2% to 6%. As depicted in Table 5, the amount of additional variance contributed varied by English language proficiency. Among third grade students, R-CBM contributed an additional 3% variance for ELLs at beginning to intermediate English proficiency and 2% variance for native English speakers. R-CBM did not contribute additional variance for third grade ELLs at early advanced to advanced English proficiency. Among fifth grade students, R-CBM contributed an additional 6% variance for ELLs at beginning to intermediate English proficiency, 3% for ELLs at early advanced to advanced English proficiency, and 3% for native English speakers. Similar to the results for research question 2a, Maze was found to be a significant predictor only for third grade ELLs at early advanced to advanced English proficiency ($R^2 = .03$).

**Research Question 3**

The third research question examined the predictive accuracy of R-CBM, Maze, and teacher judgment cut-scores using receiver operating characteristic (ROC) analyses. The outcome measure, CST-ELA-RC performance, was dichotomized using grade-specific cut-scores ($\geq .73$ for third grade and $\geq .70$ for fifth grade). The cut-scores were selected based on the CST Post-Test Guide (CDE, 2012), which indicated that these
scores represent the “average percent correct for a representative sample of the state’s students who scored at the lowest score for proficient” on the CST-ELA (e.g., scale score of 350). The CST Post-Test Guide (CDE, 2012) notes that these CST-ELA-RC cut-scores represent a practical benchmark because California Board of Education’s goal is for all students to be minimally proficient or above on the CST. These cut-scores correspond to an overall CST-ELA score of proficient or above. To assess predictive accuracy, sensitivity and specificity values were obtained for the following R-CBM and Maze cut-scores indicating “risk”: 64 (third grade R-CBM), 8 (third grade Maze), 97 (fifth grade R-CBM), and 13 (fifth grade Maze). These norm-referenced cut-scores represent performance at the 15th percentile, which indicates at-risk status for future reading difficulties. Based on recent research conducted by AIMSweb, cut-scores corresponding to a 50% probability of success on state assessments correspond approximately to scores at the 15th percentile (Pearson, 2011). Therefore, AIMSweb set their default cut-scores for risk-status at the 15th percentile. According to AIMSweb, another rationale supporting the use of the 15th percentile to indicate risk is that research indicates that approximately 15% of nation-wide students are at severe risk for reading difficulties (Pearson, 2011). Therefore, it is reasonable for AIMSweb to use that same percentile as a guide when setting its cut-scores for risk-status (Pearson, 2011). For teacher judgment, a cut-score of <4 (representing less than “proficient”) on the TRS-R was used to indicate risk. This cut-score was chosen based on logic. Since the TRS-R was based on the California State Standards, a teacher rating of less than proficient on the TRS-R would correspond with less than proficient on the CST-ELA-RC under circumstances of perfect predictive
accuracy. The effect of English language proficiency was examined by conducting separate analyses for each group (e.g.: native English speakers, ELLs at early advanced to advanced English proficiency, and ELLs at beginning to intermediate English proficiency).

The sensitivity, specificity, and area under the curve (AUC) values associated with each predictor’s cut-score are displayed in Table 6 organized by grade and English language proficiency. AUC in an indicator of overall classification accuracy and represents the probability of each predictor accurately classifying students. AUC ranges from 0.5 to 1, with 1 representing perfect classification. Overall, low sensitivity values were found for R-CBM and Maze, which ranged from .09 to .40 for third grade R-CBM, .09 to .36 for third grade Maze, .23 to .62 for fifth grade R-CBM, and .17 to .52 for fifth grade Maze. The specificity values were found to be relatively higher ranging from .84 to .96 for third grade R-CBM, .79 to .98 for third grade Maze, .84 to .95 for fifth grade R-CBM, and .66 to .94 for fifth grade Maze. In contrast to R-CBM and Maze, the TRS-R yielded higher sensitivity values (third grade ranging from .78 to .94 and fifth grade ranging from .86 to .97) and lower specificity values (third grade ranging from .12 to .47 and fifth grade ranging from .14 to .52). AUC values ranged from .57 to .79 for third grade R-CBM, .56 to .71 for third grade Maze, .64 to .77 for third grade TRS-R, .66 to .78 for fifth grade R-CBM, .55 to .64 for fifth grade Maze, and .65 to .80 for fifth grade TRS-R.

Regarding the influence of English language proficiency, AUC values were greatest for native English speakers for all grades and measures except for fifth grade
Maze. AUC values were lowest for ELLs at early advanced to advanced English proficiency for all grades and measures except for third grade TRS-R. Specificity values increased with English proficiency across all grades and measures. Specifically, native English speakers exhibited the highest specificity values, followed by ELLs at early advanced to advanced English proficiency, which was then followed by ELLs at beginning to intermediate English proficiency. In contrast, sensitivity values were highest among ELLs at beginning to intermediate English proficiency across all grades and measures. For third grade students, sensitivity was lowest among ELLs at early advanced to advanced English proficiency. For fifth grade students, sensitivity values for native English speakers and ELLs at early advanced to advanced English proficiency were relatively similar. All sensitivity values for R-CBM and Maze were below recommended values of greater than .70 (Rathvon, 2004) or .80 (Carran & Scott, 1992) representing adequate predictive accuracy of risk-status.

After examining the sensitivity, specificity, and AUC values of the cut-scores, the results were examined to locate cut scores that represent adequate levels of sensitivity for R-CBM and Maze. Based on recommendations by Johnson and colleagues (2009) and Johnson and colleagues (2010), sensitivity was set to .90 in order to minimize false negatives. The resulting cut-scores and specificity values for these analyses are displayed in Table 7. Overall, all cut-scores were increased from the AIMSweb recommended cut-scores and differed by English language proficiency. The scores ranged from 118 to 136 for third grade R-CBM, 22 to 28 for third grade Maze, 130 to 165 for fifth grade R-CBM, and 26 to 39 for fifth grade Maze. The corresponding specificity values decreased,
ranging from .23 to .42 for third grade R-CBM, .02 to .18 for third grade Maze, .22 to .35 for fifth grade R-CBM, and .06 to .10 for fifth grade Maze. For R-CBM and Maze at both grades, ELLs at beginning to intermediate English proficiency yielded the lowest cut-scores. For third grade R-CBM and fifth grade Maze, native English speakers and ELLs at early advanced to advanced English proficiency yielded the same cut-scores (e.g.: R-CBM 136 and Maze 39). For third grade Maze, ELLs at early advanced to advanced English proficiency yielded the highest cut-score (e.g.: 35). For fifth grade R-CBM, native English speakers yielded the highest cut-score (e.g.: 165).

**Research Question 4**

The effect of student reading achievement on the contribution of teacher judgment in screening was examined by conducting separate regression analyses for students with low, medium, and high reading achievement as measured by AIMSweb R-CBM and Maze. Students who scored below the 25th percentile on both measures were assigned to the low achieving group. Students who scored between the 25th and 75th percentile on at least one measure were included in the medium achieving group. The medium achieving group also included students who scored high on one measure but low on the other. Students who scored above the 75th percentile on both measures were assigned to the high achieving group. The following predictors were entered separately in the following order: 1) previous year’s CST-ELA-RC score, 2) R-CBM score, 3) Maze score, and 4) teacher judgment score. The outcome variable was the current year’s spring CST-ELA-RC score. The statistical significance of each predictor was examined following its entry into the regression model. Non-significant predictors were not retained. The proportion of
variance explained was examined following the entry of each significant predictor. See Table 8 for a summary of the proportion of variance explained by each predictor.

The following is a summary of notable findings specific to the predictive contribution of teacher judgment. Teacher judgment was found to contribute the most additional variance in predicting spring CST-ELA-RC scores for the high achieving group. It contributed an additional 7% for third graders and 9% for fifth graders. Teacher judgment was the only significant predictor for high achieving third graders. For the medium achieving group, teacher judgment contributed an additional 4% of the variance for both grades. For the low achieving group, teacher judgment contributed an additional 5% of the variance for third graders. However, it did not contribute additional variance for fifth graders. Aside from the findings regarding teacher judgment, it is interesting to note that the combination of all predictors accounted for a greater proportion variance among medium achieving students (total $R^2 = .37$ and .41) than for high achieving students (total $R^2 = .07$ and .35) or low achieving students (total $R^2 = .28$ and .18).

**Research Question 5**

The effect of teachers as a source of variability in the accuracy of their judgments was examined through multi-level modeling by comparing a random coefficient model to a more restricted fixed coefficient model. To test if random effects indicate significant between-teacher variation in slopes and intercepts, the full model (teacher judgment – reading achievement slope set as random) was compared to a restricted model (teacher judgment – reading achievement slope set as fixed). The likelihood ratio test was used to determine if the full model was a significant improvement from the restricted model. The
following conditional level-1 model and unconditional level-2 models were tested. Level 1: \( Y_{ij} = \beta_{0ij} + \beta_1 \times (\text{Prior CST}) + \beta_2 \times (\text{R-CBM}) + \beta_3 \times (\text{Maze}) + \beta_4 \times (\text{Teacher Judgment}) + r_{ij} \).

Level 2: \( \beta_{0ij} = \gamma_{00} + u_{0j}, \beta_1 = \gamma_{10}, \beta_2 = \gamma_{20}, \beta_3 = \gamma_{30}, \beta_4 = \gamma_{40} + u_{4j} \). The analyses were conducted separately for third and fifth grade. The following is a description of the model components:

- \( Y_{ij} \): This represents reading achievement as measured by spring CST-ELA-RC, which was the outcome in this study.
- \( \beta_{0ij} \): This represents the random intercept at level 1, which became the outcome variable at level 2. For level 1, it is the expected value on \( y \) (CST-ELA-RC) when \( x \) for Prior CST-ELA-RC, R-CBM, and Maze is equal to the mean (e.g.: grand mean centered) and teacher judgment is group mean centered. It was set as a random effect because the mean level of individual variables is likely to vary across teachers.
- \( \beta_1 = \gamma_{10} \): This represents the average prior year’s CST-ELA-RC – reading achievement slope. It was set as a fixed effect because there is no prior theory to suggest that it will vary at level 2 across teachers.
- \( \beta_2 = \gamma_{20} \): This represents the average R-CBM – reading achievement slope. It was set as a fixed effect because there is no prior theory to suggest that it will vary at level 2 across teachers.
- \( \beta_3 = \gamma_{30} \): This represents the average Maze – reading achievement slope. It was set as a fixed effect because there is no prior theory to suggest that it will vary at level 2 across teachers.
• $\beta_{4j} = \gamma_{40} + u_{4j}$: This represents the average teacher judgment – reading achievement slope for each teacher. It was set as a random effect because there is some literature to suggest that it may vary at level 2 across teachers.

• $r_{ij}$: This is the variance at level 1, which represents each individual’s deviation in reading achievement from their teacher means.

• $\gamma_{00}$: This is the grand mean, which represents the mean of the teacher means in reading achievement.

• $u_{0j}$: This is the variance at level 2, which represents the unique effect of teacher $j$ on mean reading achievement.

• $u_{4j}$: This is the variance at level 2, which represents the unique effect of teacher $j$ on the mean teacher judgment – reading achievement slope.

For both third and fifth grade analyses, the results of the likelihood ratio test indicated that the full random coefficient model (teacher judgment – reading achievement slope set as random) was not a significant improvement from the restricted fixed coefficient model (teacher judgment – reading achievement slope set as fixed). Specifically, when the full model deviance statistic was subtracted from the restricted model deviance statistic, the difference was less than the chi-square cut-off for significance. Therefore, the more restricted parsimonious model was kept as the final model. This indicated that teachers were not a significant source of variability in the accuracy of their judgments.
Discussion

This study focused on issues related to the use of curriculum-based measurement and teacher judgment to screen for at-risk readers. Based on the current literature, the predictive accuracy of screening tools do not meet the guidelines of 90% sensitivity as recommended in the recent screening research (Johnson et al., 2009; 2010). Therefore, there is a clear need to conduct more research in this area in order to inform improvements in the predictive validity and accuracy of screening tools. This study examined the predictive validity and accuracy of screening measures, use of teacher judgment with curriculum-based measures as a part of screening, and moderators of screening validity and accuracy. The following is a discussion of the findings in these areas.

Curriculum-Based Measures in Screening

Overall predictive validity and accuracy. In the initial validity examination of curriculum-based measures in screening, the relationship between curriculum-based measures and performance on high stakes state assessments was noted. The correlational evidence indicates that the relationship between R-CBM and future reading outcomes ranged from $r = .32$ to $r = .61$. The majority of the values were within the range of documented correlations between ORF and state standardized reading achievement tests, which have ranged from .49 to .75 (Hintze & Silberglitt, 2005; Kranzler et al., 1999; Reschley et al, 2009; Silberglitt et al., 2006, & Wood et al., 2006). The relationship between Maze and future reading outcomes was found to range from $r = .22$ to $r = .38$. 
The relatively lower correlation of Maze with future reading outcomes when compared to R-CBM has also been documented in past studies (Silberglitt et al., 2006).

In addition to the relationship between curriculum-based measures and future reading performance, this study examined the predictive contribution of curriculum-based measures (e.g.: R-CBM and Maze) in the context of screening. The results indicate that after controlling for performance on prior year’s state test, R-CBM contributed an additional 11% of the variance for third graders and 10% for fifth graders. This indicates that R-CBM serves as a useful screening tool even after considering prior year’s test performance. After controlling for performance on prior year’s state test and teacher judgment, R-CBM contributed an additional 2% of the variance for third graders and 4% for fifth graders. This indicates that although there is some overlapping variance explained by teacher judgment and R-CBM, R-CBM also provides unique predictive contribution. Maze did not contribute additional predictive variance after controlling for performance on prior year’s state test and R-CBM.

A majority of previous research examining the predictive contribution of R-CBM for future reading performance have not controlled for performance on previous year’s state assessment. Although it was not a research question in this study, the predictive contribution of R-CBM without controlling for previous year’s state assessment was examined for purposes of comparison with prior research. It was found that R-CBM contributed 32% to 35% of the variance in future reading performance, which is within the range of documented values in the literature (e.g.: 30% to 49%; Baker et al., 2008; Muyskens et al., 2009; Pearce & Gayle, 2009).
In examining the predictive accuracy of curriculum-based measures, the results suggest that the AIMSweb established cut-scores for R-CBM and Maze are relatively adequate for classifying students who are not at-risk for poor reading outcomes. Overall specificity values for third and fifth grade ranged from 92% to 95%. Despite this high level of specificity, sensitivity levels were relatively lower ranging from 28% to 36%. This indicates that these cut-scores are inadequate for classifying students who are at-risk for poor reading outcomes. These findings are consistent with the current literature that has found cut-scores to yield relatively lower sensitivity and higher specificity (e.g.: Goffreda et al., 2009; Hintze & Silberglitt, 2005; Keller-Margulis et al., 2008; Pearce & Gayle, 2009; Stage & Jacobsen, 2001). However, there are also some findings that suggest the opposite (e.g.: Hosp et al., 2001; McGlinchey & Hixson, 2004; Muyskens et al., 2009; Wood, 2006). Regarding predictive accuracy, it is also noted that positive predictive values (e.g.: proportion of correctly identified students out of all who were identified as at-risk by the screener) were greater than expected under chance prediction (e.g.: predicting all students as at-risk). For the third grade sample, the base rate for poor reading outcomes was 57%. Positive predictive values ranged from 85% to 87%. For the fifth grade sample, the base rate for poor reading outcomes was 54%. Positive predictive values ranged from 80% to 89%.

After examining the sensitivity and specificity values of the AIMSweb established cut-scores, cut-scores representing adequate levels of sensitivity were created. Based on recommendations by Johnson and colleagues (2009) and Johnson and colleagues (2010), sensitivity was set to .90 in order to minimize false negatives. In the
context of screening, it is important to minimize false negatives because the purpose of screening is to accurately identify students with poor later outcomes. One trade-off of increases sensitivity, however, is decreased specificity. Compared to the established cut-scores, all cut-scores corresponding with 90% sensitivity were higher than those established by AIMSweb.

**Role of student English language proficiency.** In the schools, the purpose of screening assessments is to identify students who are at-risk for poor later outcomes. In order to maximize the utility of screening assessments for all students, it is important to focus on predictive validity as a possible source of bias when the purpose of a test is to predict outcomes (Linn, 1973). According to educational and psychological testing standards as set forth by AERA/APA/NCME (1999), as well as other many other psychometricians (Cole & Moss, 1993; Linn, 1976), predictive bias occurs when a test makes more accurate predictions for one group than it does for another group. It occurs when a given test score is differentially valid for any subgroup of individuals who have taken the test (Cole & Moss, 1993; Linn, 1973). Although the research has yielded mixed results, English language learner status has been identified as a possible source of bias in screening assessments (e.g.: Hosp et al., 2011; Klein and Jimerson, 2005; Johnson et al., 2009; Johnson et al., 2010; Muyskens et al., 2009; Roehrig et al., 2008; Wiley & Deno, 2005). One possible explanation for these mixed results is that studies have typically characterized ELLs as a homogeneous group or, at best, as separated by native language. Level of English language proficiency has not been considered as a source of variation. However, according to recommendations for test standards set forth by AERA, APA, and
NCME (1999), language proficiency should be considered as a part of assessment and interpretation.

**Effect on predictive validity.** The results of this study provide initial evidence for the effect of language proficiency on the predictive validity of curriculum-based measures. In examining the impact of language proficiency on the relationship between R-CBM and future reading outcomes, the results for the third grade sample of this study indicate that the relationship was significantly weaker for ELLs with higher language proficiency than for native English speakers and ELLs with lower language proficiency. For the fifth grade sample, the relationship was significantly stronger for native English speakers. However, there was no significant difference in the strength of this relationship between ELLs at different language proficiency levels. According to the correlational analyses, it appears that ELL status acts as a moderator for the relationship between measures of oral reading fluency and future reading performance. In conjunction with the correlational analyses, the results from the predictive analyses suggest that the predictive contribution of R-CBM is moderated by English language proficiency. After controlling for performance on prior year’s state assessment, R-CBM contributed an additional 11% of variance for native English speakers, 7% for ELLs with higher proficiency, and 8% for ELLs with lower proficiency for the third grade sample. For the fifth grade sample, R-CBM contributed an additional 9% of variance for native English speakers, 6% for ELLs with higher proficiency, and 10% for ELLs with lower proficiency.

Although the pattern is unclear, student language proficiency appears to moderate the unique predictive contribution of curriculum-based measures of oral reading fluency.
Taken as a whole, the patterns in these results suggest that R-CBM has greater predictive validity for native English speakers than for ELLs. This may be explained by the role of English vocabulary in reading comprehension (Proctor et al., 2005). Since R-CBM is a measure of oral reading fluency, students’ English vocabulary skills may not have been captured by R-CBM. In considering the expected lower English vocabulary skills of ELLs when compared to native English speakers, the finding that R-CBM has greater predictive validity for native English speakers is not surprising.

Regarding the impact of language proficiency on the relationship between Maze and future reading outcomes, the results for both the third and fifth grade sample indicate that the correlation for native English speakers was significantly higher than that for ELLs. This relationship was non-significant for all ELLs except third grade ELLs with lower language proficiency. These findings are in contrast with the results of Wiley and Deno (2005), which found significant correlations between Maze and future reading outcomes on a state standardized assessment ($r = .73$ for third and fifth grade native English speakers and $r = .52$ to .57 for ELLs). In examining the results of the predictive analyses, it is noted that after controlling for prior year’s state assessment and R-CBM, Maze was not found to be a significant predictor of future reading outcomes except for with third grade ELLs with higher language proficiency. Taken as a whole, these results suggest that Maze has limited predictive validity in predicting future reading performance. This finding somewhat contrasts with the findings of Wiley and Deno (2005), which indicated that Maze accounted for significant variance beyond ORF in predicting performance on a high stakes reading assessment for native English speakers.
but not for ELLs. However, it is important to note that the ELLs in Wiley and Deno’s study were from Hmong rather than Spanish-speaking backgrounds. In addition, performance on prior year’s state test was not considered.

**Effect on predictive accuracy.** After noting the influence of English language proficiency on predictive validity of R-CBM and Maze, it is also important to note its influence on predictive accuracy. Some researchers have suggested that the predictive accuracy of cut-scores on curriculum-based measures may vary by language status; therefore, native English speakers and ELLs may need different cut-scores in order to maximize predictive accuracy (Jenkins et al., 2007). This present study further explored this issue by considering English language proficiency among ELLs in addition to language status. This study suggests that curriculum-based measures are relatively accurate at identifying students who are not at-risk for poor reading outcomes. Specificity levels were found to increase with English proficiency across both grades and measures. Native English speakers exhibited the highest specificity values, followed by ELLs at early advanced to advanced English proficiency, which was then followed by ELLs at beginning to intermediate English proficiency. Despite this high level of specificity, curriculum-based measures were less accurate at identifying students with negative later reading outcomes. Sensitivity values were also moderated by English language proficiency. They were highest among ELLs at beginning to intermediate English proficiency across both grades and measures.

Regarding overall correct classification accuracy, R-CBM was the most accurate in classifying native English speakers. Similar to the results of the predictive validity
analyses, these results suggest that R-CBM has greater predictive accuracy for native English speakers than for ELLs. This may again be explained by the role of English vocabulary in reading comprehension (Proctor et al., 2005), which is not adequately captured by measures of oral reading fluency. In considering the expected lower English vocabulary skills of ELLs, this finding is not surprising.

After examining the sensitivity and specificity values of the established cut-scores for curriculum-based measures, cut-scores corresponding with 90% sensitivity were created based on recommendations by Johnson and colleagues (2009) and Johnson and colleagues (2010). Compared to the established cut-scores, cut-scores corresponding with 90% sensitivity differed by English language proficiency. For R-CBM and Maze at both grades, ELLs with lower language proficiency yielded the lowest cut-scores. For third grade R-CBM and fifth grade Maze, native English speakers and ELLs with higher language proficiency yielded the same cut-scores. For third grade Maze, ELLs with higher language proficiency yielded the highest cut-score. For fifth grade R-CBM, native English speakers yielded the highest cut-score. Although the pattern is unclear, the results of this study suggest that predictive accuracy of curriculum-based measures is moderated by language proficiency.

**Theoretical connections.** The moderating effect of student English language proficiency on the predictive validity and accuracy of curriculum-based measures can be discussed in the context of reading theory. Hoover and Gough’s (1990) Simple View of Reading asserts that both decoding and listening comprehension influence reading comprehension. This relationship can be conceptualized as “reading comprehension =
decoding x listening comprehension.” According to this theory, the more fluent readers are at decoding, the more resources become available for comprehension. In general, the results of this study found curriculum-based measures to be more valid and accurate in predicting the future reading outcomes of native English speakers than for English language learners. In considering Hoover and Gough’s theory, one explanation for the moderating effect of English language proficiency is that English language learners likely have lower levels of English listening comprehension skills than native English speakers. Since reading comprehension is conceptualized as product of both efficient decoding (e.g.: reading fluency) and listening comprehension, it is not surprising that the English language proficiency was found to moderate the predictive validity and accuracy of R-CBM for future reading outcomes.

In addition to reading theory, the results of this study can also be discussed in the context of Cummins’s (1980) theories on second language proficiency. According to Cummins, Cognitive Academic Language Proficiency (CALP) is necessary for ELLs to effectively process more context-reduced and academic language present in today’s educational environment. Although Basic Interpersonal Communication Skills (BICS) often develops within two years, CALPS requires five to seven years in order to fully develop. In considering Cummins’ view on language proficiency, the lower levels of CALP occurring with ELLs (vs. native English speakers) provide a second explanation for the moderating effect of English language proficiency on the predictive validity of R-CBM for future reading performance. Thus, provided with the logical assumption that native English speakers have adequate levels of CALP, the issue of CALP does not
muddle the predictive relationship between R-CBM and future reading outcomes. However, for ELLs, perhaps level of English language proficiency (e.g.: CALP) needs to be considered in addition to R-CBM when predicting future reading outcomes.

Role of Teacher Judgment in Screening

**Overall predictive validity and accuracy.** Although the current literature has not explored the role of teacher judgment in predicting future performance on high-stakes state accountability assessments, the results of this study provide initial evidence that teacher judgment is a useful predictor worthy of consideration and further study in the screening research. The correlational evidence indicates that the relationship between teacher judgment and future reading outcomes range from $r = .27$ to $r = .60$. The majority of these values were lower than the range of documented correlations between teacher judgment and reading performance, which have ranged from .47 to .79 (Bates & Nettelbeck, 2001; Begeny et al., 2008; Feinberg & Shapiro, 2003, 2009; Madelaine & Wheldall, 2005). However, it is important to note that in these existing studies, teacher judgment is typically compared with students’ concurrent reading performance rather than future reading performance.

In addition to the relationship between teacher judgment and future reading performance, this study examined the predictive contribution of teacher judgment when used in combination with curriculum-based measures (e.g.: R-CBM and Maze) as a part of screening. Although there are a few studies that examine the effect of adding teacher judgment to other screening measures (e.g.: Flynn & Rahbar, 1998; Speece et al., 2010), the effect of combining teacher judgment with curriculum-based measures to screen for
students at-risk for poor reading outcomes as measured by performance on high-stakes state accountability assessments has not been examined. In general, the results indicate that after controlling for performance on prior year’s state test and performance on curriculum-based measures (e.g.: R-CBM and Maze), teacher judgment contributed unique variance in predicting future reading outcomes. Specifically, it contributed an additional 4% of the variance for third graders and 3% for fifth graders. When only controlling for performance on prior year’s state test, teacher judgment contributed an additional 12% of the variance for third graders and 9% for fifth graders. This indicates that although there is some overlapping variance explained by teacher judgment and curriculum-based measures, each screening tools also provides unique predictive contribution.

In examining the predictive accuracy of teacher judgment, this study suggests that teachers are relatively accurate at identifying students who are at-risk for poor reading outcomes. Both third and fifth grade teachers correctly identified approximately 90% of students with poor later reading outcomes as at-risk. Despite this high level of sensitivity, teachers were less accurate at identifying students with positive later reading outcomes. Third grade teachers correctly identified 41% of these students while fifth grade teachers correctly identified 46%.

**Moderators of teacher judgment.** After exploring the initial predictive validity and accuracy of teacher judgment, this study explored various moderators of teacher judgment including student English language proficiency, student level of achievement, and teacher effects. In addressing student language background, correlational analyses
suggest that ELL status acts as a moderator between teacher judgment and future reading performance. Among the ELLs groups, however, language proficiency did not appear to moderate this relationship. The results for the both the third and fifth grade sample of this study indicate that the correlation for native English speakers was significantly higher than those for each of the ELL language proficiency groups. However, the correlations for the ELL groups were not significantly different from each other. Although the correlational evidence suggests that language proficiency does not appear to moderate the relationship between teacher judgment and future reading performance, the results from the predictive analyses in the second research question suggest that the predictive contribution of teacher judgment after controlling for performance on prior year’s state test and curriculum-based measures is moderated by student language proficiency.

Specifically, the amount of unique variance explained by teacher judgment varied by student English language proficiency.

Among third grade students, teacher judgment contributed the greatest amount of unique variance (e.g.: 6%) for ELLs with higher language proficiency. Teacher judgment only contributed 3% unique variance for third grade native English speakers and 2% unique variance for ELLs with lower language proficiency. Among fifth graders, teacher judgment contributed the greatest amount of unique variance (e.g.: 4%) for native English speakers. It contributed an additional 2% variance for ELLs with higher language proficiency. Teacher judgment contributed no unique variance for ELLs with lower language proficiency. It appears that the lack of predictive contribution for fifth grade ELLs with lower language proficiency is due to the overlapping variance explained by
curriculum-based measures (e.g.: R-CBM) and teacher judgment. When performance on curriculum-based measures was not controlled, teacher judgment did contribute unique variance for this group. However, the primary question of interest in this study was the contribution of teacher judgment after controlling for both performance on prior year’s state test and curriculum-based measures.

Although the predictive contribution of teacher judgment increased with English language proficiency for the fifth grade sample, the pattern is less clear for third grade students. However, with both samples, teacher judgment contributed the lowest amount of unique variance for ELLs with lower language proficiency. These results suggest that after controlling for performance on prior year’s state test and curriculum-based measures English language proficiency moderates the predictive contribution of teacher judgment in predicting future reading performance. One explanation for these findings is that the predictive accuracy of teacher judgments affected their predictive contribution. The low predictive accuracy in teacher judgments for ELLs with lower language proficiency corresponds with the lower predictive contribution of teacher judgment. Specifically, teachers correctly classified only 64% of third graders and 68% of fifth graders with lower language proficiency. In contrast, teachers correctly classified 77% of third graders and 80% of fifth graders who were native English speakers.

In examining the effect of student achievement on the predictive contribution of teacher judgment for future reading performance, it was found that teacher judgment exhibited greater predictive contribution for higher achieving than for lower achieving students. For high achieving third grade students, teacher judgment was found to be the
only significant predictor for future reading performance (explaining 7% of the variance). Previous year’s test scores and reading screeners (e.g.: R-CBM and Maze) were not significant predictors for high achieving third graders. For average and low achieving third graders, teacher judgment was found to yield similar predictive contributions. After controlling for previous year’s test scores and reading screeners, teacher judgment explained an additional 4% of the variance for average achieving students and an additional 5% of the variance for low achieving students. Although the percentage of additional variance in future reading performance explained by teacher judgment for high achieving students (e.g.: 7%) appears only slightly greater than for the average and low achieving students (e.g.: 4% and 5%), it is important to note that teacher judgment was the only significant predictor for the high achieving group. Results for the fifth grade sample also indicated that teacher judgment exhibited greater predictive contribution for higher achieving than for lower achieving students. For high achieving fifth grade students, teacher judgment explained an additional 9% of the variance in future reading performance. In contrast, it only explained an additional 4% of the variance for average achieving students. For low achieving students, teacher judgment was not found to be a significant predictor of future reading performance.

Although a limited number of studies in the literature have examined the effect of student achievement on the accuracy of teacher judgment, the findings of this current study are in concordance with previous research that has suggested level of student achievement as a moderating variable for teacher judgment accuracy. Similar to the findings of this study, reviews of the literature by Hoge and Coladarci (1989) and
Coladarci (1992) both found that teachers generally made less accurate judgments for lower-performing students than they did for higher-performing students. This finding was also present in some individual studies (Begeny et al., 2008; Coladarci, 1986). Although there are also studies that did not find this moderating effect (e.g.: Bates & Nettelbeck, 2001), the majority of previous research on this issue is generally in agreement with the findings of this present study.

Regarding the extent to which teachers act as source of variability in the predictive contribution of their judgments for students’ future reading performance, the results from both the third and fifth grade sample indicate that teachers were not a significant source of variability in the accuracy of their judgments. These findings are in concordance with some of the more recent literature suggesting that teachers do not vary in the accuracy of their judgments (Bates & Nettelbeck, 2001; Begeny et al., 2008; Demaray & Elliot, 1998; Feinberg & Shapiro, 2003). However, they are in contrast to some of the earlier literature, which suggested that judgment accuracy does vary by teacher (Coladarci, 1986; 1992; Hoge & Coladarci, 1989).

**Total Predictive Contribution of Screening Assessments**

Aside from findings regarding the unique predictive validity of curriculum-based measures and teacher judgment, several other findings related to the total predictive contribution of screening assessments were noted. Across both grades, the combination of all screening assessments predicted a greater amount of variance in future reading performance for native English speakers (46% and 49%) than for ELLs (12% to 32%). This contrast was primarily due to the greater predictive contribution of previous year’s
test performance for native English speakers (32% and 37%) than for ELLs (3% to 22%). Since the predictive contribution of previous year’s test performance is lower for ELLs, it is important to consider the use of curriculum-based measures and teacher judgment in order to aid in the prediction of future reading performance. Although it was not a research question of interest in this study, level of English language proficiency was added as a final predictor for the ELL sample in order to examine if it would shed light onto some of the unexplained variance. It was found that English language proficiency did not contribute significant additional variance in predicting future reading outcomes.

**Limitations**

When interpreting the results of this study, it is important to consider its limitations. There were several main issues in this study that may limit the generalizability of the results. First, the participant selection method and demographic background should be considered. Student participants were selected using convenience sampling rather than random sampling. Regarding demographic background, it is important to note that the majority of student participants resided in socio-economically disadvantaged areas. Students from only two grades (third and fifth) were sampled. In addition, all ELL students spoke Spanish as a first language. Considering the participant selection method and demographic background, the results of this study may be limited to student populations with similar characteristics as the sample used in this study. Caution should be taken when generalizing the results to students with different demographics (e.g.: grade-level, native language, socio-economic status).
In addition to the demographics of the study participants, issues regarding data collection also present some limitations. Specifically, the order in which data were collected presents some challenges. Teacher judgment data were collected shortly after the winter screening. Therefore, it is unknown whether or not teachers used the winter screening data to inform their judgments of student reading performance. A final practical limitation of this study was that winter screening data, rather than fall screening data, were used in all the analyses. This limits the practical implications of the results because the goal of screening is early identification of at-risk students. Fall screening represents the earliest time in the school year during which students can be identified as at-risk. When interpreting the results of any study, it is important to note limitations that may affect generalizability.

**Implications**

The results of this study have several implications for educational practice and future research. In the context of educational practice and in line with best practices, practitioners should consult various sources of data when screening for students at-risk for poor reading outcomes. Based on the results of this study, it is recommended that sources of data for best practices in screening include previous year’s test scores, curriculum-based measurement scores (e.g.: R-CBM), and teacher judgment. Although Maze appears to have face validity as a measure of reading comprehension, its added utility in screening should be questioned due to the finding in this study that Maze, in general, did not contribute unique predictive various after controlling for previous year’s state assessment and R-CBM.
For best practices in screening, it is also recommended based that student variables such as ELL status, English language proficiency, and level of student achievement be considered during the screening process. This is important in light of the results of this study, which indicate variability in the predictive contribution of each data source as a function of these student variables. Specifically, educational practitioners should consider giving teacher judgment more weight for higher achieving students than for lower achieving students. Teacher judgment should also be given less weight for students with lower levels of English language proficiency. When seeking to identify students at-risk for poor reading outcomes, current AIMSweb cut-scores should be used with caution in light of the relatively low sensitivity found in this study. This study suggests that higher cut-scores are needed to yield adequate levels of sensitivity. In addition, educators should encourage educational publishers (e.g.: AIMSweb or DIBELS) to validate cut-scores with subgroups of students possessing various levels of English language proficiency. This study provides preliminary evidence that different cut-scores may be needed for ELLs than for native English speakers. Specifically, it was found that higher R-CBM cut-scores are needed for native English speakers than for ELLs.

To address the limitations and validate the results of this current study, future research on the predictive validity of curriculum-based measurement, use of teacher judgment with curriculum-based measures as a part of screening, and moderators of teacher judgment is needed. To increase the generalizability of the present studies results, future studies should consider the use of random sampling and participants from different demographic backgrounds. Future samples of participants should be expanded to include
ELLs from various language backgrounds and proficiency levels. This current study examined Spanish-speaking ELLs with either higher or lower language proficiency. In order to further explore the effect of language proficiency on the predictive validity of screening assessments, future studies should consider separating ELLs into more distinct levels of language proficiency (e.g.: low, medium, and high proficiency). Future samples of participants should also include students from different grade levels. Since prior research has suggested that indicators of reading achievement change with age (e.g.: Perfetti, 1985), further exploration on the use of curriculum-based measurement and teacher judgment as a part of screening with older students (e.g.: middle or high school) will yield more insight. In addition to expanding participant demographics, future studies should also seek to collect teacher judgment data at the same time as student screening data in order to minimize the influence of screening results on teacher judgments. Finally, the predictive validity and accuracy of fall screening scores for performance on state assessments should be examined.
References


Table 1

Descriptive Statistics for R-CBM, Maze, Teacher Judgment, and CST-ELA-RC

<table>
<thead>
<tr>
<th></th>
<th>3rd Grade</th>
<th></th>
<th></th>
<th>5th Grade</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Skewness</td>
<td>Kurtosis</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>AIMSweb R-CBM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Participants</td>
<td>99.94</td>
<td>39.89</td>
<td>-0.08</td>
<td>-0.22</td>
<td>127.08</td>
<td>39.72</td>
</tr>
<tr>
<td>Native English Speakers</td>
<td>110.84</td>
<td>39.64</td>
<td>-0.13</td>
<td>-0.06</td>
<td>141.65</td>
<td>37.21</td>
</tr>
<tr>
<td>ELLs – high proficiency</td>
<td>110.36</td>
<td>30.27</td>
<td>-0.11</td>
<td>0.04</td>
<td>121.23</td>
<td>30.19</td>
</tr>
<tr>
<td>ELLs – low proficiency</td>
<td>78.95</td>
<td>35.79</td>
<td>-0.05</td>
<td>-0.68</td>
<td>91.77</td>
<td>31.83</td>
</tr>
<tr>
<td>AIMSweb Maze</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Participants</td>
<td>15.97</td>
<td>8.72</td>
<td>0.78</td>
<td>0.25</td>
<td>21.72</td>
<td>9.74</td>
</tr>
<tr>
<td>Native English Speakers</td>
<td>18.30</td>
<td>9.17</td>
<td>0.52</td>
<td>-0.17</td>
<td>24.26</td>
<td>9.48</td>
</tr>
<tr>
<td>ELLs – high proficiency</td>
<td>17.20</td>
<td>7.64</td>
<td>1.03</td>
<td>1.03</td>
<td>21.31</td>
<td>8.94</td>
</tr>
<tr>
<td>ELLs – low proficiency</td>
<td>11.97</td>
<td>6.96</td>
<td>1.16</td>
<td>1.39</td>
<td>14.84</td>
<td>7.80</td>
</tr>
<tr>
<td>Teacher Judgment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Participants</td>
<td>3.10</td>
<td>1.12</td>
<td>-0.10</td>
<td>-0.67</td>
<td>3.21</td>
<td>1.07</td>
</tr>
<tr>
<td>Native English Speakers</td>
<td>3.39</td>
<td>1.09</td>
<td>-0.31</td>
<td>-0.48</td>
<td>3.56</td>
<td>1.01</td>
</tr>
<tr>
<td>ELLs – high proficiency</td>
<td>3.58</td>
<td>0.86</td>
<td>-0.15</td>
<td>-0.61</td>
<td>3.14</td>
<td>0.82</td>
</tr>
<tr>
<td>ELLs – low proficiency</td>
<td>2.45</td>
<td>0.96</td>
<td>0.19</td>
<td>-0.50</td>
<td>2.26</td>
<td>0.90</td>
</tr>
<tr>
<td>CST-ELA-RC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Participants</td>
<td>0.63</td>
<td>0.22</td>
<td>-0.41</td>
<td>-0.62</td>
<td>0.60</td>
<td>0.24</td>
</tr>
<tr>
<td>Native English Speakers</td>
<td>0.68</td>
<td>0.21</td>
<td>-0.62</td>
<td>-0.20</td>
<td>0.69</td>
<td>0.21</td>
</tr>
<tr>
<td>ELLs – high proficiency</td>
<td>0.75</td>
<td>0.16</td>
<td>-0.65</td>
<td>0.04</td>
<td>0.51</td>
<td>0.19</td>
</tr>
<tr>
<td>ELLs – low proficiency</td>
<td>0.50</td>
<td>0.20</td>
<td>-0.03</td>
<td>-0.76</td>
<td>0.39</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Table 2

*Correlations for R-CBM, Maze, and Teacher Judgment with Spring CST-ELA-RC*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>3rd Grade</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native English Speakers</td>
<td>.58*</td>
<td>.38*</td>
<td>.58*</td>
</tr>
<tr>
<td>ELLs – high proficiency</td>
<td>.32*</td>
<td>-.08</td>
<td>.36*</td>
</tr>
<tr>
<td>ELLs – low proficiency</td>
<td>.49*</td>
<td>.25*</td>
<td>.43*</td>
</tr>
<tr>
<td><strong>5th Grade</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native English Speakers</td>
<td>.56*</td>
<td>.22*</td>
<td>.56*</td>
</tr>
<tr>
<td>ELLs – high proficiency</td>
<td>.35*</td>
<td>.10</td>
<td>.31*</td>
</tr>
<tr>
<td>ELLs – low proficiency</td>
<td>.34*</td>
<td>.12</td>
<td>.27*</td>
</tr>
</tbody>
</table>

*significant at p < .01
Table 3

*Presence of Significant Differences Between Correlations*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>3rd Grade</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native English Speakers vs. ELLs – low proficiency</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Native English Speakers vs. ELLs – high proficiency</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>ELLs – low proficiency vs. ELLs – high proficiency</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td><strong>5th Grade</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native English Speakers vs. ELLs – low proficiency</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Native English Speakers vs. ELLs – high proficiency</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>ELLs – low proficiency vs. ELLs – high proficiency</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>
Table 4

*Research Question 2a: Proportion of Variance Explained by Each Predictor*

<table>
<thead>
<tr>
<th></th>
<th>Previous CST-ELA-RC</th>
<th>R-CBM</th>
<th>Maze</th>
<th>Teacher Judgment</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>3rd Grade</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native English Speakers</td>
<td>0.32</td>
<td>0.11</td>
<td>.</td>
<td>0.03</td>
<td>0.46</td>
</tr>
<tr>
<td>ELLs – high proficiency</td>
<td>0.04</td>
<td>0.07</td>
<td>0.03</td>
<td>0.06</td>
<td>0.20</td>
</tr>
<tr>
<td>ELLs – low proficiency</td>
<td>0.22</td>
<td>0.08</td>
<td>.</td>
<td>0.02</td>
<td>0.32</td>
</tr>
<tr>
<td><strong>5th Grade</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native English Speakers</td>
<td>0.37</td>
<td>0.09</td>
<td>.</td>
<td>0.04</td>
<td>0.49</td>
</tr>
<tr>
<td>ELLs – high proficiency</td>
<td>0.18</td>
<td>0.06</td>
<td>.</td>
<td>0.02</td>
<td>0.25</td>
</tr>
<tr>
<td>ELLs – low proficiency</td>
<td>0.03</td>
<td>0.10</td>
<td>.</td>
<td>.</td>
<td>0.12</td>
</tr>
</tbody>
</table>
Table 5

*Research Question 2b: Proportion of Variance Explained by Each Predictor*

<table>
<thead>
<tr>
<th></th>
<th>Previous CST-ELA-RC</th>
<th>Teacher Judgment</th>
<th>R-CBM</th>
<th>Maze</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd Grade</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native English Speakers</td>
<td>0.32</td>
<td>0.12</td>
<td>0.02</td>
<td>.</td>
<td>0.46</td>
</tr>
<tr>
<td>ELLs – high proficiency</td>
<td>0.04</td>
<td>0.10</td>
<td>.</td>
<td>0.03</td>
<td>0.18</td>
</tr>
<tr>
<td>ELLs – low proficiency</td>
<td>0.22</td>
<td>0.07</td>
<td>0.03</td>
<td>.</td>
<td>0.32</td>
</tr>
<tr>
<td>5th Grade</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native English Speakers</td>
<td>0.37</td>
<td>0.09</td>
<td>0.03</td>
<td>.</td>
<td>0.50</td>
</tr>
<tr>
<td>ELLs – high proficiency</td>
<td>0.18</td>
<td>0.05</td>
<td>0.03</td>
<td>.</td>
<td>0.25</td>
</tr>
<tr>
<td>ELLs – low proficiency</td>
<td>0.03</td>
<td>0.05</td>
<td>0.06</td>
<td>.</td>
<td>0.14</td>
</tr>
</tbody>
</table>
Table 6

*Sensitivity, Specificity, and AUC Values for R-CBM, Maze, and Teacher Judgment*

<table>
<thead>
<tr>
<th></th>
<th>R-CBM</th>
<th>Maze</th>
<th>Teacher Judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SN</td>
<td>SP</td>
<td>AUC</td>
</tr>
<tr>
<td>3rd Grade</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native English Speakers</td>
<td>0.26</td>
<td>0.96</td>
<td>0.79</td>
</tr>
<tr>
<td>ELLs – high proficiency</td>
<td>0.09</td>
<td>0.95</td>
<td>0.57</td>
</tr>
<tr>
<td>ELLs – low proficiency</td>
<td>0.40</td>
<td>0.84</td>
<td>0.68</td>
</tr>
<tr>
<td>5th Grade</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Native English Speakers</td>
<td>0.23</td>
<td>0.95</td>
<td>0.78</td>
</tr>
<tr>
<td>ELLs – high proficiency</td>
<td>0.26</td>
<td>0.95</td>
<td>0.66</td>
</tr>
<tr>
<td>ELLs – low proficiency</td>
<td>0.62</td>
<td>0.84</td>
<td>0.76</td>
</tr>
</tbody>
</table>

*Note.* SN = Sensitivity, SP = Specificity.
Table 7

*Cut Scores and Specificity Values Associated with 90% Sensitivity*

<table>
<thead>
<tr>
<th>Grade</th>
<th>Group</th>
<th>R-CBM</th>
<th>Maze</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cut Score</td>
<td>Specificity</td>
</tr>
<tr>
<td>3rd Grade</td>
<td>Native English Speakers</td>
<td>136</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>ELLs – high proficiency</td>
<td>136</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>ELLs – low proficiency</td>
<td>118</td>
<td>0.25</td>
</tr>
<tr>
<td>5th Grade</td>
<td>Native English Speakers</td>
<td>165</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>ELLs – high proficiency</td>
<td>155</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>ELLs – low proficiency</td>
<td>130</td>
<td>0.28</td>
</tr>
</tbody>
</table>
Table 8

*Research Question 4: Proportion of Variance Explained by Each Predictor*

<table>
<thead>
<tr>
<th></th>
<th>Previous CST-ELA-RC</th>
<th>R-CBM</th>
<th>Maze</th>
<th>Teacher Judgment</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rd Grade</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Achieving</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Medium Achieving</td>
<td>0.25</td>
<td>0.08</td>
<td>.</td>
<td>0.04</td>
<td>0.37</td>
</tr>
<tr>
<td>Low Achieving</td>
<td>0.19</td>
<td>0.04</td>
<td>.</td>
<td>0.05</td>
<td>0.28</td>
</tr>
<tr>
<td>5th Grade</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Achieving</td>
<td>0.18</td>
<td>0.08</td>
<td>.</td>
<td>0.09</td>
<td>0.35</td>
</tr>
<tr>
<td>Medium Achieving</td>
<td>0.31</td>
<td>0.06</td>
<td>.</td>
<td>0.04</td>
<td>0.41</td>
</tr>
<tr>
<td>Low Achieving</td>
<td>0.13</td>
<td>0.05</td>
<td>.</td>
<td></td>
<td>0.18</td>
</tr>
</tbody>
</table>
Appendix A

Teacher Rating Scale of Reading (TRS-R)

Please consider the following grade-level standards and performance descriptors when completing the TRS-R.

Grade 3 Reading Comprehension Standards

Students read and understand grade-level-appropriate material. They draw upon a variety of comprehension strategies as needed (e.g., generating and responding to essential questions, making predictions, comparing information from several sources).

Performance Descriptors

Far Below Basic: Students do not understand simple grade-appropriate literary and informational texts. They do not follow explicit written directions, recognize sequential steps, identify explicitly stated main events in a plot, or identify character traits based on clear text clues.

Below Basic: Students understand simple grade-appropriate literary and informational texts. They follow explicit written directions, recognize sequential steps, identify explicitly stated main events in a plot, and identify character traits based on clear text clues.

Basic: Students understand explicit aspects of grade-appropriate informational and literary text. They comprehend written directions and use details from the text to answer literal questions. They can identify the main problem and its solution in basic narrative texts and differentiate between reality and fantasy.

Proficient: Students read and understand grade-appropriate informational and literary texts. They respond accurately to questions based on literal information in the text; they use text features to locate information; they understand the main events of the plot, and they use text clues to determine character traits.

Advanced: Students read and fully understand grade-appropriate informational and literary texts. They analyze aspects of the text as a whole, such as identifying the genre of the text and making logical predictions based on information within the text. They use text clues to infer the traits of fictional characters.
Please rate each student on this list by considering the following question: How does this student’s reading comprehension skills compare to that expected by grade-level standards?

<table>
<thead>
<tr>
<th>Far Below Basic</th>
<th>Below Basic</th>
<th>Basic</th>
<th>Proficient</th>
<th>Advanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student A</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student B</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student C</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student D</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student E</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student F</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student G</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student H</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student I</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student J</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student K</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student L</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student M</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student N</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student O</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student P</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student Q</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Student R</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Teacher Rating Scale of Reading (TRS-R)

Please consider the following grade-level standards and performance descriptors when completing the TRS-R.

Grade 5 Reading Comprehension Standards

Students read and understand grade-level-appropriate material. They describe and connect the essential ideas, arguments, and perspectives of the text by using their knowledge of text structure, organization, and purpose.

Performance Descriptors

Far Below Basic: Students do not comprehend simple aspects of grade-appropriate literary and informational texts. They do not demonstrate an understanding of explicitly stated aspects of texts, such as the major topic or problem.

Below Basic: Students comprehend simple aspects of grade-appropriate literary and informational texts. They demonstrate an understanding of explicitly stated aspects of texts, such as the major topic or problem.

Basic: Students comprehend simple aspects of grade-appropriate literary and informational texts. They demonstrate an understanding of explicit aspects of texts, including the steps in a process and the author’s stated purpose.

Proficient: Students demonstrate a good understanding of grade-appropriate literary and informational texts. They grasp key ideas, including main ideas, theme, character traits, elements of plot, and purpose of text features.

Advanced: Students comprehend a wide variety of grade-appropriate literary and informational texts. They demonstrate a full understanding of the essential message of texts, draw accurate inferences, and make connections among related ideas.
Please rate each student on this list by considering the following question: How does this student’s reading comprehension skills compare to that expected by grade-level standards?

<table>
<thead>
<tr>
<th>Student</th>
<th>Far Below Basic</th>
<th>Below Basic</th>
<th>Basic</th>
<th>Proficient</th>
<th>Advanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student A</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student B</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student C</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student D</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student E</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student F</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student G</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student H</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student I</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student J</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student K</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student L</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student M</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student N</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student O</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student P</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student Q</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Student R</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
Teacher Demographics Survey (TDS)

1) What grade do you currently teach?

☐ 3rd grade
☐ 5th grade

2) Ethnicity

☐ American Indian or Alaska Native
☐ Asian or Pacific Islander
☐ Black or African American
☐ Hispanic or Latino
☐ White
☐ Other

3) How many years have you been teaching (including the current year)? ________

4) In considering all of your years teaching, what percentage of your students have been English learners?

☐ Less than 25%
☐ 25% – 50%
☐ 51% – 75%
☐ Over 75%

5) Do you possess additional certification (excluding CLAD and EL authorization required for credential or employment) to teach English learners (e.g.: Bilingual authorization, BCLAD, BCC, Bilingual Crosscultural Specialist Credential, etc…)?

☐ Yes
☐ No

6) Do you possess additional certification to teach reading (e.g.: Reading/Language Arts Specialist Credential, Reading Certificate, etc…)?

☐ Yes
☐ No