UNIVERSITY OF CALIFORNIA
RIVERSIDE

Evaluating the Predictive Validity of DIBELS Literacy Measures with Third Grade Spanish-Speaking English Language Learners

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

Doctor of Philosophy

in

Education

by

Jennifer Sun Kim

March 2012

Dissertation Committee:
Dr. Mike Vanderwood, Chairperson
Dr. Rollanda O'Connor
Dr. Greg Palardy

The Dissertation of Jennifer Sun Kim is approved:

____

____

____

Committee Chairperson

University of California, Riverside

Acknowledgements

I would like to thank my advisor, Dr. Mike Vanderwood, for his patience and guidance as he challenged me to become an independent researcher and a critical thinker. I would also like to recognize my committee members, Dr. Rollanda O’Connor and Dr. Gregory Palardy, for their support and encouragement throughout this process.

Dedication

Thank you to my husband, Sam, for always being there to lift me up and motivate me during my journey through graduate school. To my family, your words of encouragement made me believe in myself and got me to where I am today.

ABSTRACT OF THE DISSERTATION

Evaluating the Predictive Validity of DIBELS Literacy Measures with Third Grade
Spanish-Speaking English Language Learners

by

Jennifer Sun Kim

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

This study examines the predictive validity of literacy measures from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) for Spanish-speaking English language learners (ELs). Third grade EL students were screened three times during the year with DIBELS Oral Reading Fluency (DORF) and Daze. Predictive validity of the scores was examined in relation to the spring administration of the California Standards Test (CST). Data were analyzed for the entire sample as well as disaggregated by English language proficiency level. Overall results revealed that the DORF was a better predictor of CST performance than the Daze across fall, winter, and spring screenings. Although Daze was a significant predictor when examined individually, results of hierarchical regression models indicated that once DORF was accounted for, Daze did not explain significant additional variance. The contribution of each fluency measure over and above the predictability of the previous year's CST score was also examined. Results indicated that both DORF and Daze accounted for significant additional variance, however, the amount was minimal. When considering English language proficiency level, findings indicated there was not a statistically significant difference in the predictive validity of the DORF or the Daze administered in the fall to CST performance in the spring for students of varying EL level. Predictive accuracy of the DORF and Daze was also examined for each EL group, which indicated a need for additional examination of the current DIBELS cut scores when applied to this group of students. Results from this study contribute to the research suggesting that although there is potential for the use of DIBELS measures to screen Spanish-speaking ELs, further research should be conducted.

Table of Contents

List of Tables

Chapter 1: Introduction

In the United States, English Learner (EL) enrollment is at 5,346,673 students, which is a 51.0% increase since 1997-1998 (National Clearinghouse for English Language Acquisition; NCELA, 2009). In California, the EL population consists of more than 50 different primary languages, with Spanish speaking students comprising 84.6% of all ELs (NCELA, 2011). Reports from NCELA (2011) indicate that 3, 118, 404 Latino students, or 50.4% of all students, were enrolled in California schools in 2009-2010. Of those students, 1, 242, 911 were classified as ELs.

Since the passing of the No Child Left Behind Act (NCLB; 2001), the law mandates that all students must meet certain academic standards and all schools are held accountable for student outcomes. This includes students with disabilities and limited English proficient (LEP) students (U.S. Department of Education; USDOE, 2007). States are required to develop a system of high quality assessments and all public school students must participate in the assessment. The purpose is to hold schools and states accountable and to close the achievement gap by utilizing state assessments to ensure that students are meeting state academic and grade-level expectations (Maleyko & Gawlik, 2011). Unfortunately, reports from the National Assessment of Educational Progress (NAEP; 2007) indicate that 74% of fourth grade ELs in California read below basic compared to 34% of non-ELs. Thomas and Collier (1997) estimate that about 30-40% of school-aged ELs fail to reach acceptable levels of English reading by the end of elementary school.

To illustrate this ongoing discrepancy for Hispanic EL students, the NAEP (2011) report indicated that although scores for both Hispanic and White students increased, the overall achievement gap remains. In other words, although Hispanic students' scores have increased across the years, on average, White students have consistently performed better on all assessments. Therefore, despite growth in reading scores for both groups, the achievement gap between Hispanic and White students has not changed for fourth or eighth graders from 1992 to 2009. Data collected by the U.S. Department of Education in 2009 indicated a considerable proportion of Hispanic students in grades 4 (37 percent) and 8 (21 percent) were ELs. Interestingly, at grade 4 in 2009, the gap (29 points) between non-EL Hispanic students and Hispanic EL students was larger than both the gap (15 points) between White and non-EL Hispanic students, and the gap (25 points) between all Hispanic and White students.

In California, students are required to take the state mandated California Standards Test (CST) beginning in second grade. The state goal is to have all students performing at the proficient or advanced level (CDE; 2008). According to the CST Technical Report (California Department of Education; CDE; 2011b), of all third grade students (437, 450) who participated in the ELA portion of the exam, 44% scored proficient or advanced. Of the 228, 537 third grade Hispanic students who took the ELA portion of the CST, 12% scored far below basic, 21% scored below basic, 36% scored basic, 21% scored proficient, and 9% scored advanced. When examining Hispanic students who were economically disadvantaged (83% of total number of Hispanics who took CST ELA), the numbers became even more dismal, with an increased number of students performing in the far below to below basic range. The statistics demonstrate the importance of developing a system that can identify struggling EL students, as well as native English-speaking (EO) students, so that interventions can be provided early to help those students catch up to their peers.

Early Identification and Intervention

The emphasis placed on high-stake assessments has encouraged schools to use monitoring procedures to identify students who are at-risk for not reaching proficiency on the state-mandated tests. These early identification methods allow districts to intervene and improve student outcomes (Shapiro, Solari, & Petscher, 2008). In the area of beginning reading, foundational skills include phonological awareness and alphabetic principle, as well as accuracy and fluency with connected text (National Reading Panel, 2000; National Research Council, 1998). Good, Simmons, and Kame'enui (2001) state that these skills represent valid indicators toward the ultimate goal of reading by the end of third grade.

The benefit of early identification for struggling readers has been documented throughout the research. In a longitudinal study examining the development of literacy in first through fourth graders, Juel (1988) found that there is a .88 probability that a child who is a poor reader at the end of first grade will remain a poor reader at the end of fourth grade. In other words, initial differences in reading ability persist over time if no intense interventions are provided. Similarly, Stanovich (1986) found that students who fail to acquire adequate reading skills in first grade often continue to have difficulties and may never catch up to their peers. He likened this phenomenon to the "Matthew effect," in which the literacy gap increases as good readers keep improving while poor readers fall farther behind. Stanovich stated that this process occurs because better readers are exposed to more text than poor readers and are also more efficient in learning new words from context.

Chall, Jacobs, and Baldwin (1990) noted that students transition from "learning to read" to "reading to learn" in fourth grade. In other words, students begin to read for information rather than focus on learning basic reading skills, such as decoding and basic sight words. According to Robb (2002), at this point, students are expected to apply basic literacy skills gained in the earlier grades. That being said, third grade becomes an important time to intervene because many view it as the last stage where students focus on developing critical foundational skills.

Despite the potentially dismal outlook for struggling readers, many studies have shown that early intervention can help students catch up to, or even surpass, their peers. The National Reading Panel (NRP, 2000) report emphasizes that early intervention is more effective than later remediation. In fact, similar to native English speakers, research with ELs suggests that those at-risk for reading difficulties make significant progress when provided with systematic and explicit intervention in reading (Leafstedt, Richards, & Gerber, 2004; Lesaux & Siegel, 2003; Vaughn et al., 2006).

Leafstedt et al. (2004) provided a 10-week early literacy intervention for 18 kindergarten Spanish-speaking ELs. When compared to a control group that received general classroom instruction only, the authors found that the ELs in the intervention outperformed the control group, consisting of both ELs and native English speakers. The authors concluded that EL learners can meet end-of-the-year standards that are established for monolingual students. Based on a review of EL literature, Gersten et al. (2007) suggest that students with low levels of English proficiency can make comparable gains in reading when provided with supplemental reading instruction.

Response to Intervention

One challenge for educators working with a diverse population of students is the difficulty in distinguishing general reading and learning problems from reading and learning problems due to low linguistic proficiency (Geva, 2000). In the past, ELs were overrepresented in special education, indicating that perhaps their language difficulties may have been interpreted as a general learning problem (Langdon, 1989). More recently, educators appear hesitant in identifying ELs as having a learning disability and tend to wait to diagnose until the student has reached a certain level of linguistic proficiency (Limbos & Geva, 2001). The discrepancy model, otherwise known as the "wait to fail" model, can be detrimental for students who would benefit from immediate intervention (National Institute for Literacy, 2006).

Response to Intervention (RtI) is a three-tiered prevention model that focuses on early identification and evidence-based instruction (Fuchs & Fuchs, 2001). In Tier 1, all students are screened three times per year to identify those who are not meeting grade-level standards. The identified students are then provided a scientifically-based intervention, matched to their needs, as a supplement to their general education curriculum in Tier 2. Student progress is monitored and decisions are made based on the level of performance and rate of growth over time. If it is determined that a student is not making adequate progress in the intervention, the student is then considered for Tier 3 services, or special education.

Research conducted on RtI with EL learners has resulted in positive findings. Vanderwood and Nam (2007) suggested that assessment and intervention research done with EL students, within an RtI framework, can be employed with ELs. According to a report by Gersten et al. (2007), which examined past early literacy studies conducted with ELs, schools can use the same reading standards for ELs that are used with native English speakers. As districts continue to move toward RtI as a model for early identification and intervention, it is critical to identify valid assessment tools that can be used to screen and monitor students from diverse populations.

Curriculum-Based Measures (CBM)

The National Research Council (1998) indicated that assessment systems that can identify reading difficulties early and prevent later reading failure need to be in place. Good, Simmons, and Smith (1998) stated that assessment procedures are needed to (a) identify children early who are experiencing difficulty acquiring early literacy skills, (b) contribute to the effectiveness of interventions by providing ongoing feedback to teachers, parents, and students, (c) evaluate the effectiveness of interventions for individual students, (d) determine when student progress is adequate and further intervention is not necessary, (e) identify accurately children with serious learning problems, and (f) evaluate the overall effectiveness of early intervention efforts. Although published, norm-referenced tests of reading have been used to assess reading skills, there are limitations (Fuchs & Fuchs, 1999; Kame’enui & Simmons, 2001).

Marston (1989) stated that published norm-referenced tests are not intended for measuring individual student progress because they can only be used to assess gain over long periods of time, usually have only one or two forms, require extensive time to administer, and do not allow for frequent administration.

Deno (1985) proposed the use of curriculum-based measures (CBM), which involve using quick and frequently administered reading tasks that are similar in difficulty to provide a global index of reading ability. The assessments need to be valid, sensitive, and have the ability to inform instruction. CBM are a viable alternative to published tests because they consist of quick fluency measures that are sensitive to change, have many alternate forms, and can be used frequently to evaluate student progress over time (Baker & Good, 1995). Deno (2003) further stated that progress monitoring data should enable educators to formatively evaluate programs and adjust instruction when they appear to be unsuccessful in helping students pass highstakes tests. He went on to state that, “CBM data has become the basis for making judgments about whether students will achieve mandated levels of performance on benchmark tests.”

The validity of CBM reading as a measure of overall reading proficiency, including comprehension, has been well-established (i.e., Deno, Mirkin, & Chiang, 1982; Shinn, Good, Knutson, Tilly, & Collins, 1992). Research has found oral reading CBM (R-CBM) to be a valid and reliable indicator of overall reading and reading comprehension (i.e., Deno, 1985; Deno, Mirkin, Chiang, & Lowry, 1980; Fuchs, Fuchs, Hosp, & Jenkins, 2001), and predictive of later reading ability (i.e., Fuchs et al., 2001; Shinn, 1998). Correlations between R-CBM and standardized measures of reading comprehension range from .63 to .90, with most coefficients above .80 (Marston, 1989). Furthermore, many studies have found high correlations (.49-.81) between reading CBM scores and performance on high-stakes assessments across many states (i.e., Crawford, Tindal, and Stieber, 2001; McGlinchey and Hixon, 2001; Shapiro, Keller, Lutz, Santoro, and Hintze, 2006; Wood, 2006). Studies have shown that with CBM, instructional quality and student achievement increase (i.e., Fuchs, Deno, & Mirkin, 1984; Fuchs, Fuchs, Hamlett, & Stecker, 1990). In fact, CBM is frequently cited as a potential method for improving the quality of services within both the regular and special education settings (i.e., Christenson, Ysseldyke, & Thurlow, 1989; Gersten, Carnine, & Woodward, 1987).

Baker and Good (1995) examined the reliability, validity, and sensitivity of using reading fluency in English with native Spanish-speaking second grade students who were bilingual in Spanish and English. The Stanford Diagnostic Reading Test (SDRT) and teacher ratings of reading ability served as the outcomes. Grade-level oral reading fluency measures were randomly selected from the curriculum and were administered twice a week for 10 weeks. Results showed that English reading fluency was as reliable and sensitive to reading growth for ELs as for native English speakers. The authors concluded that oral reading fluency is a valid index for reading proficiency, including reading comprehension, for both ELs and English-only students.

Chapter 2: Review of Selected Literature

Oral Reading Fluency and Statewide Reading Assessments

Oral reading fluency (ORF) involves the development of multiple components that work together to result in proficient reading (Adams, 1990, Fuchs et al 2001, Wolf & Katzir-Cohen, 2001). In this paper, reading fluency will reflect the ability to orally read connected text with speed and accuracy (Adams, 1990; Kame'enui & Simmons, 2001). When assessing oral reading fluency, passages may be taken from a basal reader, however, ORF is generally assessed with CBM, which measures the number of words read correctly in 1-min (Deno, 1985; Shinn, 1998).

ORF is a commonly used tool to assess reading progress and predict later reading outcomes in second through sixth grade (Shapiro et al., 2008). Many school districts are using ORF assessments to assist with instructional decisions and intervention planning for students who are at-risk for unsuccessful performance on statewide tests (i.e., Good, Simmons, & Kame'enui, 2001; McGlinchey & Hixon, 2004; Stage & Jacobson, 2001). As with all measurement tools, concurrent and predictive validity of the assessment tools must be established before they can be considered an effective tool for prediction.

Stage and Jacobson (2001) investigated the predictive validity of ORF to performance on the state-mandated Washington Assessment of Student Learning (WASL). Fourth grade students were given ORF measures in the fall, winter, and spring, and the WASL was administered in the spring. The CBM passages were developed using grade-level reading passages from the school's reading curriculum. The students in the study were predominately European American. The authors found that ORF at all three time points reliably predicted WASL reading performance, with a medium effect size for each correlation (*r* = .43, .43, .44, respectively). The level of ORF performance in the fall, winter, and spring was a better predictor of performance on the WASL than growth in ORF across the school year (*r* = .26), indicating that slope was not a significant predictor of the WASL.

Crawford, Tindal, and Stieber (2001) used ORF in second and third grade to predict third grade performance on statewide achievement tests in reading and math. Fifty-one students were assessed in both second and third grade. ORF measures were composed of three passages from the Houghton Mifflin Basal Reading Series. Students were tested with the CBM measures one time in January, during both years of the study. Students were administered the Oregon state test in March of third grade. The correlations between second and third grade ORF with third grade reading test scores were .66 and .60, respectively. The correlation between second and third grade ORF performance was .84.

McGlinchey and Hixon (2001) examined the correlation and predictive ability of ORF and performance on the statewide Michigan Educational Assessment Program (MEAP) across eight years. Fourth grade general and special education students were included in the study. Three randomly selected passages from the fourth grade district reading basal served as the ORF passages. The ORF passages were administered two weeks prior to the MEAP. Results indicated a moderately strong relationship between ORF and the state reading assessment, ranging between .49 to .81 across the years. Some limitations of the study included different assessment times across the different years, the proximity of the testing windows between ORF and the MEAP, and the fact that only a correlation analysis was conducted to analyze predictability. In addition, this study did not differentiate between native English speakers and EL students.

As part of their study, Wiley and Deno (2005) examined the predictive validity of ORF for third and fifth grade Hmong, Somali, and Spanish ELs, as well as non-EL students. ORF was measured every two weeks from November to May, using a triad of *Standard Reading Passages*. Maze assessments were also administered in the fall, winter, and spring. The state assessment, the Minnesota Comprehensive Assessment (MCA), was administered in March. Results indicated significant correlations between ORF and the MCA for both ELs and non-ELs, in both third and fifth grade. In third grade, correlations between ORF and the MCA were slightly higher for non-ELs (*r* = .71) than for ELs (*r* = .61). In fifth grade, correlations between ORF and the MCA were higher for ELs (*r* = .69) than non-ELs (*r* = .57). An implication of this study was that the maze appeared to be a better predictor than ORF for non-EL students in fifth-grade, and significantly contributed to performance on the MCA for both third and fifth grade non-EL students. However, ORF appeared to be slightly more predictive of MCA performance than the maze for EL students in both third and fifth grade.

Wood (2006) examined the relationship between ORF and performance on the Colorado statewide reading test, the Colorado Student Assessment Program (CSAP). Participants included third, fourth, and fifth grade students whose primary language was English. Students were administered the winter benchmark of ORF two months prior to the CSAP. Pearson's *r* correlation between ORF and the CSAP score was statistically significant for each grade: third grade (*r* = .70), fourth grade (*r* = .67), and fifth grade (*r* = .75). Results indicated that oral reading fluency added unique information to predicting CSAP performance, over and above the predictability of previous year CSAP testing for fourth and fifth grade students. CSAP testing begins in third grade so there was no prior grade CSAP data for third grade students. The correlation between third grade CSAP and 4th grade CSAP was .73. When third grade CSAP and fourth grade oral reading fluency were entered simultaneously, both were significant and independent predictors of fourth grade CSAP. When ORF was entered into equation after third grade CSAP, it accounted for 9% of the variance in fourth grade CSAP. Together, third grade CSAP and fourth grade ORF accounted for 62% of the variation across CSAP scores.

The correlation between fourth grade CSAP and fifth grade CSAP was .81. When entered simultaneously, both were significant and independent predictors of fifth grade CSAP. When ORF was entered into the equation after fourth grade CSAP, it accounted for an addition 4% of variance. Together, fourth grade CSAP and fifth grade ORF accounted for 70% of the variation in fifth grade CSAP scores. Results of this study are consistent with previous studies indicating that ORF is strongly correlated with performance on statewide reading assessments (Crawford et al., 2001; Fuchs et al, 2001; Good et al., 2001; McGlinchey & Hixon, 2004; Stage & Jacobson, 2001). Woods (2006) concluded that although ORF accounted for only a small amount of variance after prior year CSAP performance was included, the finding that ORF is able to predict performance beyond previous year CSAP performance is promising for identifying needs, instructional planning, and intervention for students at different reading levels. A limitation to this study is that the measures were given within two months of each other, which does not sufficiently allow for early intervention.

As part of their study, Shapiro, Keller, Lutz, Santoro, and Hintze (2006) investigated the relationship between AIMSweb ORF passages and performance on the state assessment, the Pennsylvania System of School Assessment (PSSA), for third, fourth, and fifth-grade students in two districts. ORF was administered in October, February, and May, while the PSSA was administered in March/April. Results indicated that that all correlations were statistically significant, ranging between .62-.69, except for the fall assessment in the second district (r = .25). Overall, the results indicated that the ORF measures had a moderate to strong relationship with the state assessment for both third and fifth grade students. Winter and spring ORF data contributed the most to PSSA outcomes.

Reschly, Busch, Betts, Deno, and Long (2009) conducted a meta-analysis examining the correlational evidence between oral reading fluency and standardized measures of reading for students in grades 1-6. Results indicated that oral reading fluency scores serve as good indicators for how well students are likely to perform on reading achievement tests (weighted average r = .67). Oral reading fluency was more strongly correlated with individually administered achievement tests than group administered tests. Although results indicated that oral reading fluency was more strongly correlated with national group administered tests than state-specific tests, the strength of association between oral reading fluency and state tests was significant and moderately strong. The authors concluded that using oral reading fluency for screening and identifying students who may be at risk for low performance in state tests is warranted, however, other reliable and valid measures should be used to support high-stakes decisions.

Although most studies have shown positive results when using ORF, many teachers remain skeptical that ORF results provide an accurate indicator of reading comprehension (Fuchs, Fuchs, and Maxwell, 1988). Some teachers believe that many students can read fluently but do not understand what they read. Hamilton and Shinn (2003) found that a common concern is that ORF only measures decoding skills and not general reading ability or comprehension. This is an issue because if teachers do not view information as meaningful or important, they are less likely to use the data to inform their instruction (Allinder & Oats, 1997).

Hamilton and Shinn (2003) investigated this concept of "word callers" to see if ORF overestimated the ability of students who could decode but not comprehend. Findings indicated that students identified as "word callers" had significantly lower ORF scores than students whom teachers estimated to read well. Fuchs, Fuchs, Hamlett, and Ferguson (1992) found that despite evidence against the idea of "word callers," many teachers believe the maze, which is another general outcome measure (GOM) that has been used to assess reading growth, is a better measure of comprehension because students need to understand what they are reading in order to select the correct answers.

Maze as an Indicator of Overall Reading Ability

Fuchs et al. (1992) described the maze task as a global measure of reading, because it requires decoding, fluency, and comprehension. Maze passages are typically designed with every *n*th word replaced with three multiple choice answers. Students are asked to read the passage silently and circle the correct word. Although the maze was originally designed as untimed (Cranney, 1972), they became timed measures to parallel oral reading measures (Deno, Maruyama, Espin, & Cohen, 1990). Parker, Hasbrouck and Tindal (1992) found that timed scores are less negatively skewed and likely to increase validity coefficients.

Early studies on maze procedures have shown strong correlations between the maze and standardized reading measures. Guthrie, Siefert, Burnham, and Caplan (1974) reported a correlation of .82 between performance on the maze and standardized tests, with retest reliability over .90. Espin, Deno, Maruyama, and Cohen (1989) also examined the relationship between performance on the maze and a standardized measure of reading, and found a strong correlation between the number of correct words read and the number correct on the maze (*r* = .86) (as cited in Fuchs & Fuchs, 1992).

Fuchs and Fuchs (1992) assessed the criterion validity of four reading measures, including Question Answering Tests, Recall Procedures, Cloze Techniques, and Maze Procedures. Question Answering involves a teacher formulating and asking questions related to text and evaluating the accuracy of response. Recall Procedures required a student to read passages and retell what occurred in the passage using their own words. The Cloze Technique omitted every *n*th word and the student was asked to replace the omitted word with a meaningful word. The Maze Procedure was a modified cloze task where every *n*th word was replaced with three choices and students were asked to select the word that meaningfully fits the blank. The maze was given in a timed format. Results indicated that the cloze and retell methods were inadequate for monitoring student growth. However, the validity of the maze was strong and technically similar to ORF. Additionally, teacher satisfaction with the maze was high and teachers used the maze to design more effective instructional programs. Fuchs and Fuchs also found that the stability of maze data were higher than that of other reading measures for students with mild disabilities.

Jenkins and Jewell (1993) investigated the relationship between R-CBM, maze, teacher judgment, and standardized achievement tests. Participants include 335 subjects in grades 2-6. Teachers were asked to rank the lowest 15 students in their class. Students were then administered the Gates-MacGinitie Reading Tests (MacGinitie, Kamons, Kowalski, MacGinitie, & McKay, 1978) during the second week of school, the Metropolitan Achievement Tests (MAT; Prescott et al., 1984) during the third week of April, and three maze passages from the Basic Academic Skills Samples (BASS; Espin et al., 1989) and three oral reading passages developed by Deno, Marston, Deno, and Marston (1988), during the third week of school, in April, and in May. The authors found a significant relationship between all the measures.

Results indicated that R-CBM correlated higher with total reading achievement and comprehension than the maze. Correlations ranged between .65-.76 between the maze and Gates total reading scores, and ranged from .66-.76 between the maze and the MAT total reading. Correlations between oral reading fluency and Gates total reading ranged from .67-.88, while the correlation between oral reading fluency and the MAT total reading ranged from .60-.87. The authors found a decrease in the relationship between R-CBM and the achievement measures as grade level increased. Whereas correlations between oral reading fluency and the Gates total reading were .83, .88, and .86 in grades 2, 3, and 4, respectively, the correlation declined to .67 in sixth-grade. This study indicated that although teachers may perceive the maze as a better measure of reading comprehension than oral reading fluency, R-CBM was more correlated with teacher rankings of their students. The authors concluded that both the maze and oral reading fluency showed a strong relationship to reading achievement scores, but grade-level may affect the concurrent validity of reading fluency measures. The correlation between the maze and reading achievement did not show a decline as grade-level went up. Correlations ranged between the mid 60s and 70s with no trend based on grade-level. Based on the data, the authors also suggested that the maze was not as sensitive for less skilled readers, and appeared more sensitive to differences at upper grade levels. Oral reading fluency and maze performance correlated with teacher judgments at .66 and .56, respectively. An important note is that this study used reading passages around the second-grade difficulty level for all students.

Ardoin et al. (2004) examined the validity of using R-CBM and maze for universal screening. Participants included 77 third grade students, who were predominantly White and African American. R-CBM and maze passages were selected from the Silver, Burdett, and Ginn (1991) reading series. Students were also administered the Letter-Word Identification (LWI), Reading Fluency (RF), and Passage Comprehension (PC) subtests from the WJ-III Achievement test (Woodcock, McGrew, & Mather, 2001), which yielded a Broad Reading (BR) scale. In addition, students were administered the Iowa Test of Basic Skills (ITBS: Hoover, Hieronymus, Frisbie, &

Dunbar, 1996) in the spring. The WJ-III, maze, and R-CBM probes were administered within 5 days of each other, whereas the ITBS was administered 10 weeks after. Correlations between R-CBM and the WJ-III subtests, the maze and WJ-III subtests, and R-CBM and the maze, were statistically significant. Correlations between the median R-CBM score and the WJ-III BR, RF, and LWI were significantly higher than correlations between the maze and these subtests.

However, the correlation between the median R-CBM score and the WJ-III PC was not significantly greater than the correlation between the maze and the WJ-III PC. The WJ-III BR scores were strongly correlated with performance on both R-CBM (*r* = .70) and the maze (*r* = .50). Using hierarchical multiple regression, the authors found that the addition of the maze score did not explain significant unique variance in the WJ-III BR beyond the predictive value of R-CBM. Similarly, maze did not contribute as a predictor of the WJ-III PC after R-CBM was accounted for. The ITBS was highly correlated with both the WJ-III BR (r = .68) and R-CBM (*r* = .64). R-CBM and the ITBS were relatively equal predictors of the WJ-III BR, but each measure explained significant additional variance in the WJ-III BR that was not accounted for by the other measures. Based on the results, Ardoin et al. concluded that R-CBM appeared to be a better predictor of both comprehension and total reading ability than the maze. The authors stated that when conducting universal screenings, the maze would not explain a significant amount of variance beyond that of the R-CBM. In fact, when entered simultaneously into the regression models, only R-CBM remained a significant predictor of both comprehension and total reading achievement. In addition, R-CBM may be a better predictor of reading achievement than some norm referenced achievement tests.

As part of their study, Wiley and Deno (2005) explored the predictive validity of the Maze for third and fifth grade native English speakers, and Hmong, Somali, and Spanish ELs. Students were screened in the fall, winter, and spring using maze passages from the Basic Academic Skill samples (BASS; Deno et al., 1990). They were also administered oral reading fluency passages every two weeks. The state-mandated Minnesota Comprehensive Assessment (MCA) was administered in March. Results indicated that there were significant correlations between the Maze and the MCA. This was true for both ELs and non-ELs, in both third and fifth grade. In third grade, correlations between the Maze and the MCA were slightly higher for non-ELs (*r* = .73) than for ELs (*r* = .52). In fifth-grade, correlations between the Maze and the MCA were again slightly higher for non-ELs (*r* = .73) than ELs (*r* = .57).

Findings from this study support the use of the Maze as a tool for screening and progress monitoring the reading skills of ELs. When the maze was entered after oral reading scores in a multiple regression analysis, results indicated that the maze accounted for significant variance in MCA scores beyond oral reading for non-EL students in third and fifth grade. It did not account for additional variance for EL students. When maze was entered first and oral reading was entered second, marginal variance was accounted for non-EL third grade students. An important implication of this study is that Maze measures are predictive of performance on the MCA. Although oral reading fluency appears to be slightly more predictive of MCA performance than the maze for EL students in both third and fifth grade, the Maze appears to be a better predictor than oral reading fluency for non-EL students in fifth grade. It also significantly contributes to performance on the MCA for both third and fifth grade non-EL students.

Due to the strong correlations between ORF and the maze task with overall reading ability, these measures have often been used to identify students at-risk for reading difficulties and to monitor progress in overall reading skills. Carnine (1997) states that a valid assessment system should be used during the primary grades to document whether the students are learning enough. The system should allow for reasonable and reliable predictions of whether students who perform well are likely to perform at benchmark, or grade-level expectations, in the following years.

Dynamic Indicators of Basic Early Literacy Skills (DIBELS)

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) is a set of reliable and valid early literacy measures that can be used to assess reading skills (Good & Kaminski, 2002). DIBELS measures are general outcome measures (GOM) that utilize generic measurement procedures with stimulus materials that are not drawn from the curriculum (Shinn, 1998). The measures are not intended to diagnose but to provide indications of academic performance in relation to early reading skills (Good et al., 1998). The measures are simple to administer, inexpensive in terms of time and resources, and sensitive to improvement in skills over time.

Goffreda et al. (2009) point out that unlike curriculum-based measurements that are sampled from a district's curriculum, the DIBELS consists of standardized items. The measures are considered dynamic because pre-reading skills are assessed on a continual basis using indicators that represent the key elements of basic early literacy skill (Good and Kaminski, 2002).

DIBELS measures are criterion-referenced assessments and each measure has an empirically established benchmark goal that changes across time to ensure that skills are developing in a manner predictive of continued progress. Benchmark goals were developed by longitudinally tracking a group of students and comparing performance levels on critical early literacy skills in kindergarten and first-grade to future reading performance (Good et al., 2001). The comparisons allowed for predictions about which students were progressing adequately and which students may have needed additional instructional support. Based on performance, students are classified into risk categories with related instructional recommendations, which allow educators to identify at-risk students and provide an appropriate reading intervention (Good & Kaminski, 2002).

The DIBELS measures adhere to the five big ideas in English literacy instruction highlighted by the National Reading Panel (2000), which include phonological awareness, phonics, vocabulary, fluency, and comprehension. The DIBELS provide different timed fluency measures that assess each skill, and many studies have been conducted on these assessments to examine the psychometric properties of these fluency measures. The current study focuses specifically on fluency and comprehension.

Shaw and Shaw (2002) examined the utility of the DIBELS Oral Reading Fluency (DORF) assessment in predicting performance on the Colorado state assessment for third grade students. A total of 52 students were administered the DORF in the September, January, and April of the 2001-2002 school year. The Colorado State Assessment

Program (CSAP) was administered in April 2002. Results indicated that spring DORF was highly correlated with the CSAP (*r* = .80). There was also a relatively strong correlation between the fall DORF and CSAP (*r* = .73) and the winter DORF and CSAP (*r* = .73). The correlation among the three DORF administrations ranged between .89-.93.

Buck and Torgesen (2003) conducted a study on the predictive validity of ORF to the Florida Comprehensive Assessment Test (FCAT) for third grade students. ORF scores were obtained in May 2002 and the FCAT was administered in April 2002. There was a significant correlation between ORF and the reading portion of the FCAT scores. In fact, the authors found that the correlations between ORF and the FCAT were similar across racial/ethnic groups (*r* = .70 for white students, *r* = .62 for African American students, *r* = .78 for Hispanic students). It is important to note that the study was conducted with predominately white students, and of the minority groups, only 1% was considered LEP. Although findings did not significantly differ across ethnic groups, due to the small percentage of minority students in the study, the authors indicated that non-significant differences may change with a larger, more diverse minority sample.

Riedel (2007) examined the relationship between the DIBELS measures, reading comprehension, and vocabulary with predominantly African American first grade students. The students were assessed with Letter Naming Fluency (LNF), Phoneme Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), Oral Reading Fluency (DORF), and Retell Fluency, as well as the Group Reading Assessment and Diagnostic Evaluation (GRA + DE) and the TerraNova. A separate correlation analysis was conducted for the few Spanish-speaking EL students included in the study. Riedel found that NWF administered in the beginning of the year was a better predictor of end-year comprehension for both ELs and non ELs (*r* = .41 and .45, respectively) than PSF (*r* = .31 and .26) or LNF (*r* = .15 and .44). Once DORF was administered, beginning in the winter of first grade, it became the best single predictor of comprehension at the end of first grade for both ELs and non ELs (*r* = .72 and .59, respectively). Although the sample of ELs in the study was small, results indicated that the correlation between DORF and comprehension was similar to the correlation for non-ELs in the study.

Roehrig, Petscher, Nettles, Hudson, and Torgesen (2008) examined the validity of the DORF in predicting performance on the Florida Comprehensive Assessment Test (FCAT-SSS) and the Stanford Achievement Test (SAT-10) reading comprehension measures. Participants included 35, 207 third grade students enrolled in Florida Reading First schools. Students were predominately White, African American, and Latino. The DORF was administered four times during the year in September, December, February/March, and April/May. The FCAT-SSS and SAT-10 were administered in February/March. Moderate to strong correlations were found among DORF scores, and results indicated that the relationship between the DORF and FCAT-SSS, and the DORF and SAT-10, increased over time, with the strongest magnitude when the measures were administered concurrently (*r* = .70-.71). The DORF predicted reading comprehension performance equally on the FCAT-SSS and the SAT-10, indicating that the DORF is able to predict performance on a state-developed measure, as well as a common measure of reading comprehension. Results also indicated that there were no significant EL effects. DORF was equally able to identify at-risk students, regardless of demographic characteristics. The most significant predictor of risk on the FCAT-SSS was DORF performance, and interactions between race, SES, and language status with DORF were not significant contributors. In other words, Roehrig et al. (2008) found that there was no bias in predictions for EL students. A limitation was that the study examined ELs as one group. If the group was disaggregated into different ethnicities and language proficiency levels, results may have been different.

Baker et al. (2008) investigated the role of DORF as a predictor of reading proficiency for students in first through third grade. The authors' objectives were to examine the relationship between DORF and a high-stakes reading test, examine whether the slope on DORF predicted performance on the high-stakes reading tests over and above initial level of DORF performance alone, and test how well DORF stood up in prediction models that included a comprehensive measure of reading. Scores on the DORF administered in the fall, winter, and spring, were compared to performance on the Stanford Achievement Test-Tenth Edition (SAT-10) and the Oregon Statewide Reading Assessment (OSRA), which were administered in the spring of third grade. Findings indicated that correlations between the DORF administered in first, second, and third grade and the SAT-10 and the OSRA assessments ranged between .58 to .82. Of the participants, 32% were classified as ELs, with Hispanics representing 68%. The authors found that ORF slope added to the accuracy of predicting performance on the OSRA in Year 2, above information provided by level of performance alone. Together, ORF level and slope explained 70% of the variance on the SAT-10 at the end of second grade. ORF level and slope accounted for 51% of the variance on the OSRA given in third grade.

Shapiro et al. (2008) examined the predictive value of oral reading fluency administered in the fall and winter to the Pennsylvania System of School Assessment (PSSA). There were 1,000 participants in third, fourth, and fifth grade. The sample consisted of White (48%), Black (34.2%), and Hispanic (11.7%) students. In addition to the DORF administered in the fall and winter, participants were given a group-administered comprehension measure, the *4Sight Benchmark Assessment*, in the fall and winter. The PSSA was administered in late winter. Results indicated there was a very strong relationship between fall and winter DORF across all grades (r = .92-.94), and a strong relationship between fall and winter 4Sight (*r* = .72-.80). There were moderate correlations between DORF and 4Sight. Results indicated that 4Sight was more correlated with PSSA than the DORF across all grade levels. The relationship between 4Sight and PSSA was stronger than the relationship between DORF and PSSA during both fall and winter assessments in third and fourth grade. There were no significant differences in fifth grade. In examining the diagnostic validity of DORF and 4Sight, the authors found that both were predictive for students who scored at or above benchmark on both of the measures and scored proficient on the PSSA. However, the predictive validity of the DORF and 4Sight was lower for students who scored below proficient on the PSSA. Across the grades, a combination of the DORF and 4Sight resulted in better classification rages. Data indicated that DORF is able to identify pass/fail rates for students who are low-risk or at-risk, but does not do well with students who are identified as some-risk on DORF.

Goffreda et al. (2009) investigated the predictive validity of scores on first grade DIBELS measures for predicting performance on two standardized reading measures in second and third grade. Participants included 67 first-grade students, of which 78% were White, 10% were Hispanic, 2% were Black, 1% was Asian, and 9% belonged to an unknown ethnic group. Eleven percent of students were in special education. Students were administered the first grade DIBELS benchmark assessments in the September, January, and May, while the TerraNova was administered in the spring of second grade, and the Pennsylvania System of School Assessment (PSSA) was given in the spring of third grade. Results indicated the correlation between DORF and TerraNova was .39, while the correlation between DORF and PSSA was .54.

Although there have been many studies examining the relationship between ORF and overall reading ability, less have been conducted with maze procedures. In addition, research done specifically with EL students has been lacking. Thus far, results have been promising for both EL and non-EL students, but studies that have included minority students have examined them as one homogenous group, without differentiation of primary language or English language proficiency level. This is a major limitation in the research base because although there is strong evidence for the use of ORF and maze tasks, results can lead to an overgeneralization to groups of students who have not been included in samples, or have been combined into a group of students with different backgrounds.

Testing Standards

The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999) state that assessment tools with test norms based on native English speakers should not be used with individuals whose first language is not English. Using assessments with individuals who have not sufficiently acquired the language of the test may result in scores that do not accurately reflect the intended construct to be measured. For example, a test designed to measure academic achievement might actually become a measure of language proficiency when used with a non-native speaker. As a result, the samples should be representative of the linguistic and cultural characteristics of the intended examinees in each stage of test development: test design, validation, and norming. Equal attention should be allocated to each linguistic group.

It is important to consider within-linguistic group differences as well when interpreting assessment results and making decisions in education programs (AERA, APA, & NCME, 1999). Students with the same linguistic background often vary in proficiency level, ranging from those who have no knowledge of the language to those who are fluent and knowledgeable of the language. Combining all ELs into one group can result in overlooking potential differences within the group (Artiles, Rueda, Salazar, & Higareda, 2005). Therefore, individual differences should be taken into account when interpreting assessments. The Standards emphasize that it is important to understand that poor test performance may result from poor language proficiency rather than other deficiencies.

A review of the literature examining the relationship between English oral proficiency and English word-level literacy skills indicates that for younger students, English language proficiency explains only a small amount of unique variance, whereas phonological processing skills, such as phonological awareness, RAN, and working memory are more consistent predictors of English word and pseudoword reading (Lesaux, Geva, Koda, Siegel, & Shanahan, 2008). Some studies have found that English oral language proficiency is associated with reading comprehension skill for ELs. Oral vocabulary knowledge is important for English reading proficiency (e.g., Jimenez, Garcia, & Pearson, 1996). Research suggests that limited vocabulary is related to low levels of reading comprehension in English, and ELs who have a large vocabulary are able to process written text better (Lesaux et al., 2008). The relationship between English proficiency and word reading skills appears to be influenced by the method used to assess proficiency, which may contribute to the lack of relationship between proficiency and word reading skills.

Predictive Bias

It has been well established that the purpose of a screening measure is to predict future outcomes (Betts et al., 2008). This is commonly examined in studies of predictive validity, which provides an indication of how well performance on a criterion measure is predicted by performance on a screening measure (Hosp, Hosp, & Dole, 2011). One area of predictive validity that is investigated less frequently is the idea of prediction bias, or differential prediction. Prediction bias is present when there is a difference in the quality of inferences when making a judgment about individuals from one group than another (Helms, 2006). In other words, bias in predictive validity occurs when a test differentially predicts an outcome for one group over another. There have been few studies that have examined the idea of bias in predictive validity with screening measures that include a span of more than 3 months between predictor and outcome variables (Betts et al., 2008).

In their study examining the accuracy of DORF in predicting third grade reading outcomes, Roehrig et al. (2008) investigated the idea of prediction bias for students from different categories of socio-economic status, language status, and race/ethnicity. In this study, 12% of the students were classified as ELs, with no indication of ethnic breakdown or English language proficiency level. Using a logistic regression analysis, the authors found no significant interaction effects, and concluded that DORF was able to equally identify at risk readers regardless of demographic characteristics, including EL status.

Similar to findings from the Roehrig et al. (2008) study, Betts et al. (2008) did not find evidence of predictive bias between EL and native English speakers when examining fluency measures of phonemic awareness, alphabetic principle, and oral reading at the end of kindergarten as predictors of reading achievement at the end of second grade. It is important to note that the EL sample comprised an almost equal number of Hmong and Hispanic students, and a lower number of Somali students, combined as one group. The authors suggested that it is possible that findings may differ with students with different native languages and different sample size ratios. For example, Spanish speaking ELs in this study scored the lowest average score on both early literacy measures and the second grade reading achievement test. Therefore, districts with higher proportions of Spanish speaking ELs may result in different outcomes.

As part of their study with first through third grade students, Hosp et al. (2011) posed the question of how much the prediction accuracy of DIBELS NWF and ORF to a state-criterion referenced test would vary as a function of the level of performance when examined across disaggregated categories of economic disadvantage, limited English proficiency, disability status, and race/ethnicity. The authors found that in general, there was less bias in predictive validity in grade 3 than grades 1 and 2. Hosp et al. stated that although there were differences in performance of different groups across measures, the patterns of potential bias were not extreme or consistent.

It is important to note that the results of previous studies often examined bias between a group native English speakers and a group of EL students. The majority of studies clustered all EL students as one homogeneous group and did not mention ethnic background or English proficiency levels. As discussed in the Standards, it is important to consider across group and within group differences during test validation.

Predictive Accuracy

When investigating the predictive validity of a measure, it is also important to examine the accuracy of classification. The purpose of screening is to identify students who are on the trajectory of failure (Johnson, Jenkins, Petscher, & Catts, 2009). A screening measure needs to be accurate in distinguishing between students who will have future difficulties from those who will not. Diagnostic accuracy involves the examination of the cutoff scores on a criterion measure that is used to classify students as being at-risk for failure (Rathvon, 2004). There are four possible outcomes of a screening measure: valid positive, valid negative, false positive, and false negative (see Table 1). Students who are identified correctly fall in the valid categories, while students who are identified incorrectly fall in the false categories. Valid positive rates refer to the number of students who are correctly identified as being at risk. Valid negative rates reflect the number of students who are correctly identified as not being at risk. The false positive category indicates the number of students that are identified as being at risk, but are not truly at risk. The false negative category refers to the number of students who are identified as not at risk, but are truly at risk.

The main indices of diagnostic accuracy are sensitivity, specificity, positive predictive power, and negative predictive power (Rathvon, 2004). Sensitivity reflects the ability of the screener (early literacy measures) to correctly identify students who become poor readers. It is calculated by comparing the number of valid positives with the number of students who develop reading difficulties. Specificity reflects the ability of the screener to correctly identify students who do not become poor readers and is calculated by comparing the number of valid negatives with the number of students who later become adequate readers. When sensitivity levels decrease, more students who are truly at risk will be missed (Johnson et al., 2009). As specificity levels decrease, more students are identified as at risk when they are not really at risk.

According to Jenkins (2003), the minimum level for sensitivity should be high, possibly as high as 90%, so that students who are truly at risk are identified and given targeted instruction. Other researchers have suggested that sensitivity, specificity, and positive predictive power should be between 75-80% (i.e., Carran & Scott, 1992; Kingslake, 1983; Rathvon, 2004). In terms of the acceptability of false positive rates, Catts el al. (2009) suggest that a rate of 50% or less seems acceptable.

Positive predictive power refers to the proportion of valid positives among all students the screener identifies as at-risk for reading difficulties (Rathvon, 2004). Negative predictive power refers to the proportion of valid negatives among all students the screener identified as not at-risk and who later become adequate readers. The hit rate reflects the overall effectiveness of the screener.

Accuracy of cut-scores is important because of the implications it can have on intervention time and educational resources (O'Connor & Jenkins, 1999). If the cut-scores over identify the number of students that are at-risk, or mistakenly identify students as at-risk for reading difficulties, this can tax limited educational resources (Johnson et al., 2009). Alternatively, if cut-scores under-predict risk status, students who need additional instruction will miss out on valuable intervention time. This can be detrimental for these students, as research has clearly demonstrated the benefits of early intervention for struggling readers (i.e., Leafstedt et al., 2004; Lesaux & Siegel, 2003; Vaughn et al., 2006).

Past research has found that DIBELS measures are moderately accurate in predicting which students will read well or struggle at the end of first grade (Fuchs, Fuchs, & Compton, 2004; Riedel, 2007). When examining the diagnostic accuracy of Letter Naming Fluency (LNF), Phoneme Segmentation Fluency (PSF), and Nonsense Word Fluency (NWF), Riedel (2007) found sensitivity levels between 61-68%, and specificity levels between 60-68%, with PSF demonstrating the lowest levels of both. This indicates that the DIBELS measures missed 32-39% of truly at risk students and over identified 32-40% of students.

In their study on the predictive validity of ORF to the Florida Comprehensive Assessment Test (FCAT) for third grade students, Buck and Torgesen (2003) found that ORF scores predicted FCAT-SSS reading scores with 92% specificity and 77% sensitivity. Shapiro et al. (2008) investigated that diagnostic efficiency of the DORF benchmark scores to the PSSA with predominately White and Black students in grades 3, 4, and 5. DORF screening measures were administered in the fall and winter, as was the 4Sight benchmark assessment. The authors found a sensitivity level of 96% and a specificity level of 55% for fall administration of the DORF for third grade students. During the winter of third grade, the sensitivity and specificity levels for DORF were 95% and 59%, respectively.

As part of their study examining the utility of DORF in predicting performance on the Colorado state assessment for third grade students, Shaw and Shaw (2002) reported an overall hit-rate of 74% in appropriately predicting CSAP performance levels. The specificity of classifying low risk students was 90%, but the sensitivity of identifying at-risk students was lower (43%). The authors noted that lowering the cutoff score from 110, which was the DIBELS benchmark goal for Grade 3, to 90, which was the goal for Grade 2, resulted in better classification accuracy. The overall hit-rate using the Grade 2 benchmark goal increased the accuracy of student classification to 86%, with specificity and sensitivity levels increasing to 90% and 73%, respectively.

Hintze, Ryan, and Stoner (2003) examined the concurrent validity and diagnostic accuracy of the DIBELS with the Comprehensive Test of Phonological Awareness (CTOPP; Wagner, Torgesen, & Rashotte, 1999). Eighty-six kindergarten students were administered DIBELS LNF, ISF, and PSF, as well as the CTOPP in one session, with the order being counterbalanced. Results indicated that DIBELS measures strongly correlated with most subtest and composite scores on the CTOPP. ISF and PSF cut-scores based on the DIBELS criteria resulted in very high levels of sensitivity and low levels of specificity, meaning that these measures tend to over-identify students as at-risk. The percentage of true positives was high, but there were also a large number of false positives. The authors suggest that if this result is typical, the measures should only be used for screening, and thereafter, identified positive students should be re-evaluated with another measure. However, if the measures are intended to be used for decisions, then lower cut-scores need to be determined.

Nelson (2008) investigated the diagnostic accuracy of DIBELS ISF, LNF, PSF, and NWF in determining the risk status of kindergarten students with the Test of Phonological Awareness (TOPA-2) and the Woodcock-Johnson Tests of Achievement (WJ-III). The DIBELS measures and TOPA-2 were administered in January and the WJ-III was administered in May. Results indicated that PSF and NWF cutoff scores for determining at-risk status showed higher sensitivity than ISF and LNF. For some-risk status, ISF, PSF, and NWF had sensitivity indexes of 80-90% and false positive rates of 41-73%. Sensitivity rates for LNF were 53-72%. Overall, DIBELS cutoff scores for at-risk status showed low levels of sensitivity. Depending on the measures, the authors indicate that up to 68% of the at-risk students were missed. Cutoff scores for some-risk had higher sensitivity but lower specificity, indicating that there were higher false positive rates. In addition, the DIBELS measures showed low positive predictive power but high negative predictive power, which indicates that the measures were better at identifying students with adequate reading skills than identifying those with inadequate reading skills. The authors conclude that although moderate to strong correlations were found between DIBELS and the criterion measures, the overall diagnostic accuracy showed that the utility of the measures was only moderate.

Wood (2006) examined the predictive accuracy of the DORF and the Colorado state test (CSAP). He found that the oral reading fluency scores were effective in predicting which students would pass or fail the CSAP. Both positive predictive power and negative predictive power were above the base rate of failure for grades 3, 4, and 5. For third grade students, sensitive was 86% and specificity was 64%.

Vanderwood, Linklater, and Healy (2008) examined the predictive accuracy of NWF for first-grade EL students and found that the DIBELS criterion for NWF was adequate in identifying ELs who may need additional instruction. NWF scores were able to correctly identify almost 80% of the students who performed above the 25th percentile on all of the outcome measures in third-grade. However, NWF scores were not as accurate in predicting who would score below the 25th percentile on the outcome measures. In addition, the authors found that over 80% of the false negatives in the study comprised students with the lowest English proficiency. This suggests that additional information needs to be considered when determining the at-risk classification of students with the lowest levels of language proficiency. Overall, findings provide initial support for using NWF to screen and identify ELs who may need additional services, but it may not be as accurate for students with the lowest level of English proficiency. If language proficiency level had not been examined in this study, this difference would not have been identified.

Despite recommendations by the Standards (AERA, APA, & NCME, 1999), most studies that have examined the utility of early literacy measures in identifying, predicting, and progress monitoring students who are at-risk for reading difficulties have primarily been conducted with native English-speaking students (e.g., Kaminski & Good, 1996). As the number of ELs from different cultural backgrounds continues to rise, it is important that early literacy measures be validated with these subgroups as well.

As more school districts move toward using screening data to identify and group students for intervention, rather than relying on test scores from the previous year, there is a need for additional studies to examine the predictive validity of timed fluency measures with specific EL subgroups. Furthermore, the idea of prediction bias relating to EL students of differing English language proficiency levels is warranted to identify potential differences based upon group membership. In addition, current cut scores should be examined for their utility with students of different ethnicities and EL levels.

Research Questions

The current study investigated the predictive validity of the DORF and Daze with third grade Spanish-speaking ELs. The specific research questions are:

1. What is the relationship amongst the predictor variables, DORF, Daze, CELDT, and CST 2010, and between the predictor variables and the outcome variable, CST 2011?
2. To what extent does the relationship between DORF and CST 2011, and Daze and CST 2011, differ across the fall, winter, and spring screenings for students of varying English language proficiency levels?
3. To what extent is performance on the DORF and the Daze, administered in the fall, predictive of performance on the CST 2011, administered in the spring?
4. To what extent does fall performance on the DORF and Daze predict performance on the CST 2011, above and beyond the predictability of the previous year's CST score?
5. To what extent does prediction bias play a role in the predictive validity of the DORF and the Daze, administered in the fall, for students of varying English language proficiency levels?
6. What is the diagnostic accuracy of the DIBELS cut scores for DORF and Daze, in the fall, winter, and spring, for third grade students of differing language proficiency levels, as measured by sensitivity, specificity, positive predictive power, and negative predictive power?

Chapter 3: Methods

Participants

A total of 607 Spanish speaking EL students were assessed, however 85 cases were removed from the analysis due to missing data and/or unknown English language proficiency level. The final sample included 522 third grade Spanish EL students (301 males and 306 females) from six elementary schools. Based on CELDT scores, a breakdown of EL levels were as follows: 34 Beginners, 88 Early Intermediates, 291 Intermediates, 94 Early Advanced, and 15 Advanced. Student ethnicities at each school ranged between 97-99% Hispanic, followed by <1% African American, <1% Asian, <1% White, and <1% Other. Across the schools, 79-89% of students were classified as ELs, with 99-100% listing Spanish as the home language.

Due to the small sample size at the beginning and advanced levels of English language proficiency, the end groups were combined to create a more meaningful comparison among groups. Students at the beginning (B) and early intermediate (EI) levels on the CELDT represented one group, the B/EI group (N = 122). Students at the early advanced (EA) and advanced (A) levels on the CELDT were combined to represent the EA/A group (N = 109). Students who scored at the intermediate level on the CELDT formed their own group, the Int group (N = 291).

Setting

The Southern California school district participating in this study has been gradually moving toward district-wide RtI implementation. RtI implementation began during the 2008-2009 school year, with ten field study schools at the first grade level. RtI extended to all elementary schools at the first grade level, and to the second grade at the original ten field study schools during the 2009-2010 school year. During the 2010-2011 school year, RtI implementation expanded to all elementary schools at the first and second grade level. Six schools implemented up to the third grade level, while three implemented RtI schoolwide.

The data for this study were collected from the six elementary schools implementing RtI at the third grade level. The district served predominately lower socioeconomic families, with 88-98% of the students receiving free or reduced lunch. There were a total of 23 third grade classrooms, with an average of 4 classrooms per school. There were 21 female teachers and 2 male teachers. The language of instruction in these classrooms was English, regardless of English language proficiency level. All students received two hours daily of Open Court Reading, which was the district adopted reading program. Every EL student received 30 minutes of English Language Development (ELD) instruction per day using the Carousel of IDEAS program (Ballard & Tighe, 1976). This comprehensive and systematic English language development program was designed for K-5 EL students at all levels of language proficiency. The program integrates listening, speaking, reading, and writing through interactive activities that emphasize fine literature, phonics, and development of literacy skills (Ballard & Tighe, 1976).

Materials

California English Language Development Test (CELDT). The CELDT is an annual assessment of English language proficiency that is administered to students in kindergarten through twelfth grade. It assesses the skill areas of Listening/Speaking, Reading, and Writing. The purpose of the CELDT is to identify LEP students, determine the level of their English language proficiency, and to monitor their progress in areas of listening, speaking, reading, and writing in English. Schools are required to administer the CELDT to students whose primary language is not English (California Department of Education, 2010). There are 5 levels of overall performance: Beginning, Early Intermediate, Intermediate, Early Advanced, and Advanced. Students in the beginning level of proficiency demonstrate little to no receptive or productive English skills. In the early intermediate stage, students are still developing receptive and productive English skills. They are able to identify and understand more concrete details during instruction and may be able to communicate with more ease, with frequent errors.

Students who perform in the intermediate level are beginning to tailor English skills to communicate and learn. They are able to respond with increasing ease to more varied communication and learning demands. Students in the early advanced stage are able to use English to learn in academic situations. At the advanced level, students can communicate effectively and are able to use English in complex and demanding situations with infrequent errors. Students are classified as fluent English proficient (FEP) if their overall score is at the early advanced level or above and scores in each domain are in the intermediate category or above. Students are classified as EL if their overall score is below the early advanced level or one of the domain scores is below the intermediate level. According to the CELDT technical manual, the reliability coefficients for the CELDT are between .71 to .91 across all grades and domains (CDE, 2011a). Overall scores between 230-414 are in the beginning range, 415-459 are in the early intermediate range, 460-513 are in the intermediate range, 514-556 are in the early advanced range, and 557-700 are in the advanced range (CDE, 2011a).

DIBELS Next Oral Reading Fluency (DORF). DORF is a standardized, individually administered test of accuracy and fluency with connected text. Students are asked to read 3 passages aloud for 1 min each. Words omitted or substituted, and hesitations of more than 3 sec, are scored as errors. Words that are self-corrected within 3 sec are score as correct. The median number of words per minute (wpm) the student reads from the 3 passages is the oral reading fluency rate. Test-retest reliabilities for elementary students range from .92-.97. Alternate-form reliability of different reading passages drawn from the same level ranges from .89 to .94 (Tindal, Marston, & Deno, 1983). Criterion-related validity studied in eight separate studies in the 1980s reported coefficients ranging from .52 to .91 (Good & Jefferson, 1998). For ELs consisting of Hmong, Somali, and Spanish, correlations between DORF in the third grade and the Minnesota Comprehensive Assessment (MCA) was .71 (Wiley & Deno, 2005). The correlation between ORF in the fifth grade and the MCA was .57.

In the fall of third grade, students who score 55wpm or below are considered well below benchmark, students who score between 56-69wpm are considered below benchmark, and students who score 70wpm or higher are considered at or above benchmark. In the winter of third grade, 68wpm or below is considered well below benchmark, 69-85wpm is considered below benchmark, and 86wpm or above is considered at or above benchmark (see Table 2). In the spring of third grade, students who score 80wpm or below are considered well below benchmark, 81-99wpm is considered below benchmark, and 100wpm or above is considered at or above benchmark. These benchmark scores were updated from the previous edition, DIBELS 6th edition, and are based upon data collected during the 2009-2010 school year (Dynamic Measurement Group, 2010). Students who score at or above benchmark are *likely to need core support*, and have 80-90% odds of achieving subsequent literacy goals. Students below benchmark are *likely to need strategic support*, and have 40-60% odds of meeting the next literacy goal, without extra support. Students in the well below benchmark range are *likely to need intensive support*, with 10-20% odds of achieving subsequent literacy goals without intensive support.

DIBELS Next Daze. Daze is a standardized measure of reading comprehension that assesses a student's ability to construct meaning from text using word recognition skills, background information and prior knowledge, familiarity with linguistic properties such as syntax and morphology, and reasoning skills (Good et al., 2011a). Daze can be given individually, to a small-group of students, or to a whole class at one time. Approximately every seventh word in the Daze passages is replaced by a box containing the correct word and two distracter words. Students are asked to read a passage silently and to circle their word choices. An adjusted score, which compensates for guessing, is calculated based on the number of correct and incorrect responses. The two-week alternate form reliability for Daze in third grade was .75 (Good, Kaminski, Dewey, Wallin, Powell-Smith, & Latimer, 2011). Inter-rater reliability for Daze was .99. The criterion-related validity of the Group Reading Assessment and Diagnostic Evaluation

(GRADE) and third grade Daze in the fall, winter, and spring were .67, .61, and .67, respectively. Predictive validity of third grade Daze administered in the fall was .79 with winter DIBELS Composite Score, and .74 with DIBELS Composite at the end of the year (Good et al., 2011).

In the fall of third grade, students who score 5 or below are considered well below benchmark, students who score between 6-7 are considered below benchmark, and students who score 8 or higher are considered at or above benchmark. In the winter of third grade, 7 or below is considered well below benchmark, 8-10 is considered below benchmark, and 11 or above is considered at or above benchmark (see Table 1). In the spring of third grade, students who score 14 or below are considered well below benchmark, 15-18 is considered below benchmark, and 19 or above is considered at or above benchmark.

California Standards Test (CST). The CST is part of the Standardized Testing and Reporting (STAR) Program that was developed to measure and evaluate student achievement of the content standards (CDE, 2011b). The CSTs assess content standards for English Language Arts (ELA), mathematics, history-social science, and science in grades two through eleven. Scores are reported through scaled scores (150-600), which are equated with five performance levels: far below basic (150-261), below basic (262-299), basic (300-349), proficient (350-401), and advanced (402-600).

Each grade-level CST was administered to approximately 400,000-500,000 students in 2010. EL students are required to participate in state testing regardless of the length of time they have been in the U.S. or their English proficiency level. In third grade, the ELA cluster of the CST is made up of Word Analysis and Vocabulary Development, Reading Comprehension, Literary Response and Analysis, Written Conventions, and Writing Strategies. The internal consistency of the CST ELA in grade 3 was .93, and validity ranged from .79-.80. Across the 2002-2004 school years, the reliability ranged from .93-.94. Convergent validity with the CAT/6 Survey, a norm-referenced test that assesses students in reading, language, spelling, mathematics, and science, was high.

Procedures

Students were screened with grade-level DORF and Daze from DIBELS *Next* three times during the year. The measures were administered in the fall (September), winter (January), and spring (March). The DORF was administered to each student individually, while the Daze was given through whole class administration. Students who were absent during the whole group administration were given the Daze individually upon return. The CST was administered whole class in the spring of 2011. All assessments were administered in English by the classroom teachers.

Prior to administration of the assessments, teachers were provided with a 7 hour professional development training specific to DIBELS administration and scoring. During training, teachers were exposed to the materials and were given multiple opportunities to practice administration and scoring. In addition, approximately half the teachers received an additional 30 min of training at their school site by the principal investigator. No formal reliability checks were conducted.

CELDT data were obtained for each student from the current school year to determine English language proficiency level. Scores from the previous year's CST administration were also obtained in order to examine the added predictability of the DORF and Daze above and beyond that of the previous year's CST scores.

It is important to note that teachers were instructed to provide supplemental intervention to students who scored in the intensive or strategic need for support range based on the DIBELS benchmark goals after each screening. The most commonly used interventions included the Peer Assisted Learning Strategies for Grades 2-6 (PALS; Fuchs et al., 2008) and The Six-Minute Solution: A Reading Fluency Program (Adams & Brown, 2007). Teachers attended 6 hour professional development trainings specific to each intervention. On-site demonstrations and support were provided per teacher or administrator request.

Chapter 4: Results

The data were examined and the assumptions of independence, normality, linearity, homoskedasticity, and multicollinearity were assessed. To test the assumptions of linearity, independence, normality, and homoskedasticity, an examination of descriptive statistics and a visual analysis of residual plots was conducted. The points appeared to be normally distributed about the fitted line, suggesting that the assumptions were met. Skewness, kurtosis, and the presence of outliers were also examined. The data were slightly skewed to the right for fall Daze for students in the B/EI group; however, regression assumptions were still met. Based on Wetherill's (1986) recommendation, the variance inflation factor (VIF) was examined for each predictor variable, following the suggested value of VIF < 10. The VIF values for all independent variables were below the recommended value indicating that the assumption of homogeneity of variance was met.

Descriptive Statistics of Entire Sample

Means and standard deviations calculated for the entire sample are presented in Table 3. Overall, students showed a steady increase in the number of words read correctly and the number correct on the comprehension fluency measure. In the fall of third grade, students read an average of 69.81 (*SD* = 27.04) words correct per minute on DORF, which is in the strategic need for support range based on DIBELS benchmark goals. Students scored an average of 6.53 (*SD* = 4.27) on the Daze, which is also in the strategic need for support range. By the winter of third grade, the average for all students increased to 79.64 (*SD* = 27.12) words read correctly on the DORF and 9.48 (*SD* = 5.44) correct on the Daze, which both fall within the strategic need for support range. In the spring, the students read an average of 90.14 (*SD* = 31.39) words correct on DORF, which is in the strategic need for support range, and scored an average of 14.73 (*SD* = 6.69) on the Daze, which is in the intensive need for support range.

The average performance on the previous year's CST (CST 2010) was in the Basic range (*M* = 334.31, *SD* = 46.33). Average performance on the outcome measure, CST 2011, was slightly lower, but also fell within the Basic range (*M* = 307.86, *SD* = 43.56).

Descriptive Statistics Disaggregated by English Language Proficiency Level

Descriptive statistics disaggregated by English language proficiency level are displayed in Table 3. For students in the B/EI English language proficiency group, the average words read correctly on the DORF was consistently lower than the average of the entire sample and fell within the intensive need for support range at each time point, based on the DIBELS benchmark goals. In the fall, students read an average of 52.09 (*SD* = 25.76) correct words per minute on DORF. The number of words read increased at each successive time point but continued to fall within the intensive need for support range. In the winter, the B/EI students read an average of 61.67 (*SD* = 26.83) words per minute. In the spring, the average fluency rate increased to 70.64 (*SD* = 32.25) words per minute. Students at the Intermediate language proficiency level performed in the strategic need for support range during the fall (*M* = 69.46, *SD* = 22.69), winter (*M* = 79.55, *SD* = 22.82), and spring (*M* = 91.28, *SD* = 28.16) DORF screenings, with average scores similar to the entire sample. The EA/A students consistently read more words correctly per minute than the B/EI and Intermediate students. The average performance fell within the core need for support range during all screenings.

The same pattern seen with the DORF screenings was seen with the Daze screenings. The EA/A students consistently performed better at each time point than the Intermediate students, and the Intermediate students in turn scored higher than the B/EI students during each Daze screening period. B/EI students showed consistent growth from the fall (*M* = 3.93, *SD* = 3.53), winter (*M* = 6.20, *SD* = 4.29), to spring (*M* = 9.92, *SD* = 6.63) screenings. Students in the Intermediate group also showed steady growth similar to that of the overall sample in the fall (*M* = 6.63, *SD* = 4.05), winter (*M* = 9.36, *SD* = 4.99), and spring (*M* = 14.98, *SD* = 5.83). Of each group, the EA/A group displayed the most growth from each screening to the next: fall (*M* = 8.76, *SD* = 4.12), winter (*M* = 12.90, *SD* = 5.53), spring (*M* = 19.22, *SD* = 5.46).

On the previous year's CST (CST 2010), the average score for B/EI students was 291.10 (*SD* = 34.65), for Intermediate students was 333.80 (*SD* = 33.96), and for EA/A students was 380.07 (*SD* = 40.59). These scores fell in the below basic, basic, and proficient range, respectively. A similar pattern was seen for the average CST 2011 performance, where B/EI students performed in the below basic range (*M* = 262.56, *SD* = 33.42), Intermediate students performed in the basic range (*M* = 307.49, *SD* = 31.46), and EA/A students performed in the proficient range (*M* = 350.55, *SD* = 36.53).

Research Question 1: Aggregate Correlations

Aggregated correlations were conducted to answer the first research question: What is the relationship amongst the predictor variables, DORF, Daze, CELDT, and CST

2010, and between the predictor variables and the outcome variable, CST 2011? According to Cohen (1992), correlations up to .29 are considered small, correlations between .30 to .49 are considered medium, and correlations .50 and above are considered large.

Pearson correlations among variables were conducted for the entire sample to provide a reference for comparison when examining correlational results for different EL subgroups (see Table 4). A moderate to strong, positive correlation was found among all the independent and dependent variables (*p* <.01). When examining the relationships between the fluency measures (DORF and Daze) and the CST 2011, the Daze administered in the spring had the strongest correlation with the outcome measure (*r* = .58; *p* <.01). Daze administered in the fall had the weakest correlation with the outcome measure; however, the relationship was still significant (*r* = .39; *p* <.01). The relationship between DORF, across screening periods, and CST 2011 ranged between .53-.55 (*p* <.01).

Fall, winter, and spring DORF scores were strongly correlated with each other, ranging from .84-.89. The correlation among fall, winter, and spring Daze scores were lower but still significant, ranging between .48-.50. The correlation between DORF and Daze performance was significant at all time points, ranging between .53-.66, with the strongest correlation occurring between DORF Winter and Daze Spring.

The students’ English language proficiency level, as measured by the CELDT, was significantly correlated with all the independent measures, as well as the outcome measure. There was a strong correlation between CELDT and DORF Fall, CELDT and

DORF Winter, and CELDT and Daze Spring (*r* = .51, .51, and .51, respectively; *p* <.01). There was a moderate correlation between CELDT and DORF Spring, CELDT and Daze Fall, and CELDT and Daze Winter (*r* = .48, .43, and .43, respectively; *p* <.01). CELDT had a strong, positive correlation with both the previous year's CST performance (CST 2010), as well as the outcome measure, CST 2011 (*r* = .67 and .69, respectively; *p* <.01).

CST score from the previous year (CST 2010) was significantly correlated with all measures. CST 2010 had a strong correlation with both the DORF and Daze measures during each screening, ranging between .53 to .58, except for Daze Fall and Daze Winter, which indicated moderate correlations of .46 and .41, respectively. CST 2010 was highly correlated with CELDT (*r* = .67; *p* <.01) and CST 2011 (*r* = .67; *p* <.01).

The dependent measure, CST 2011, was significantly correlated with all predictor variables. The strongest correlation was between CST 2011 and CELDT (*r* = .69; *p* <.01). CST 2011 was strongly correlated with DORF Fall, DORF Winter, DORF Spring, and Daze Spring, and was moderately correlated with Daze Fall and Daze Winter. CST 2011 had a strong, positive correlation with CST 2010 (*r* = .67; *p* <.01).

Research Question 2: Disaggregated Correlations

Correlations disaggregated by language proficiency were conducted to examine if there is a difference in the relationship between the predictor variables amongst each other and with the outcome measure for students at the B/EI, Intermediate, and EA/A English language proficiency levels. This section also addresses the question: To what extent does the relationship between DORF and CST 2011, and Daze and CST 2011, differ across the fall, winter, and spring screenings for students of varying English language proficiency levels?

For the B/EI group, DORF Spring, Daze Spring, and CST 2011 were significantly correlated with all other measures, ranging between .27-.90 (p < .01), with the strongest correlation occurring between winter and spring DORF (see Table 5). The smallest, significant correlation occurred between spring Daze and CELDT. The correlations between CELDT and DORF Fall (r = .13), CELDT and Daze Fall (r = .09), CELDT and Daze Winter (r = .12), and CELDT and CST 2010 (r = .08) were non-significant. There was a small correlation between CELDT and DORF Winter (r = .19; p <.05), Daze Fall and CST 2010 (r = .24; p <.05), and Daze Winter and CST 2010 (r = .23; p <.05).

As seen in Table 6, all correlations for the Intermediate group were significant at the p <.01 level, ranging between .19-.86 for all measures, except for the correlation between Daze Winter and CST 2011 (r = .13; p <.05). The strongest correlation was evident between DORF Fall and DORF Winter (r = .86; p <.01).

For the EA/A group, most correlations were significant, ranging between .26-.86 (p <.01) (see Table 7). The strongest correlation was seen between DORF Winter and DORF Spring, while the weakest correlation was between Daze Winter and CST 2010. There was a small correlation between Daze Fall and Daze Winter (r = .26; p <.05), Daze Fall and CELDT (r = .20; p <.05), and Daze Winter and CST 2011 (r = .24; p <.05). The correlation between Daze Fall and CST 2011 was non-significant.

In response to the second research question, an examination of the relationship between DORF and CST 2011, as well as Daze and CST 2011, across different language proficiency groups, and at each screening period, was conducted. When considering the entire sample, DORF had a large, significant relationship with CST 2011 across all time points. The relationship was the strongest in the winter (*r* = .55), but similar in the fall and spring (*r* = .54 and .53, respectively). The relationship between Daze Spring and CST 2011 for all students indicated a large significant relationship (*r* = .58), whereas the relationship between Daze Fall and CST 2011 (*r* = .39), and Daze Winter and CST 2011 (*r* = .41), only showed a moderate yet still significant correlation.

For students in the B/EI group, the relationship between DORF and CST 2011 was strong and significant at all time points (*p* < .01). The correlation between the measures was strongest in the fall (*r* = .59); however, the winter and spring correlations were similar (*r* = .53 and .55, respectively). The correlation for Daze and CST 2011 was moderate during the fall (*r* = .35) and winter (*r* = .46) screenings, but was strong in the spring (*r* = .64). In fact, the relationship between Daze Spring and CST 2011 was stronger than the correlation between CST 2011 and any other measure, including CELDT and the previous year's CST scores.

For students in the Intermediate group, the relationship between DORF and CST 2011 was significant (*p* <.01) during each screening, and showed a slight increase over time. The strength of association between DORF and CST 2011 was moderate in the fall, winter, and spring (*r* = .31, .37, and .38, respectively). In terms of the correlation between Daze and CST 2011, there was a moderate, significant relationship between the two measures in the fall and spring (*r* = .30 and .41, respectively). The strength of association between the Daze Winter and CST 2011 was small but significant (*r* = .13; *p* < .05).

The relationship between the DORF and CST 2011 for students in the EA/A group was similar to that of the Intermediate group. Although the correlation was strongest in the middle of the year, the strength of association was similar across the three time points (r = .36, .39, and .36, respectively; p <.01). The relationship between Daze and CST 2011 increased over time. There was a non-significant correlation between Daze and CST 2011 in the fall (r = .15), a small, significant relationship between Daze and CST 2011 in the winter (r = .24; p < .05), and a moderately significant correlation in the spring (r = .36; p < .01).

Research Question 3: Predictive Validity

The following section addresses the question: To what extent is performance on the DORF and the Daze, administered in the fall, predictive of performance on the CST 2011, administered in the spring? Simple linear regressions, as well as hierarchical regressions, were conducted to determine the contribution of each measure, as well as the amount of variance explained when both measures were entered as predictors.

Simple Linear Regression. Regression analyses were conducted to determine the predictive validity of each fall fluency measure on the CST 2011. Cohen (1988) suggests that as a general rule, R^2 of .02, .13, and .26 can serve as operational definitions of a small, medium, and large effect size, respectively. Results are presented in Table 8. The first regression model was: CST 2011 = $B_0 + B_1$ DORF Fall. DORF Fall was significant, indicating that it is a good predictor of performance on the CST 2011 in the spring. Specifically, DORF Fall was able to explain approximately 28.8% of the variance in CST 2011. The following regression equation was yielded: $Y' = 243.23 + .91X$, indicating that for a one word increase on the DORF, CST 2011 score is expected to increase by .91 points.

Another regression model was run to determine the predictive validity of Daze Fall: CST 2011 = B_0 + B_1 Daze Fall. Results indicated that Daze Fall was a significant predictor and explained 15% of the variance in CST 2011. The model yielded the following regression equation: Y' = 282.14 + 4.19*X*, indicating that for a one point increase on Daze Fall, performance on CST 2011 was expected to increase by 4.19 points.

Hierarchical Regression. In addition to examining the predictive validity of each fluency measure on its own, hierarchical regression analyses were conducted to determine the amount of additional variance explained by each successive fluency measure in the fall to performance on the CST 2011 in the spring (see Table 9). In the first model, DORF Fall was entered in step 1 and was found to be statistically significant, explaining 32% of the variance in CST 2011. When Daze Fall was entered in step 2, there was a non-significant change in R^2 [$F_{(2, 385)} = 93.44$, $p = .064$, $\Delta R^2 = .01$].

In a separate analysis, Daze Fall was entered in step 1 and was found to be a significant predictor, explaining 15.3% of the variance in CST 2011. DORF Fall was entered in step 2 and explained an additional 17.4% of the variance. Once DORF Fall was entered in the model, Daze Fall became non-significant. The addition of DORF Fall in step 2 resulted in a significant change in R^2 [$F_{(2, 385)} = 93.44$, $p <.001$, $\Delta R^2 = .17$]. In the final model, only DORF Fall was significant (t = 9.96, $p < .001$). Altogether, Daze Fall and DORF Fall explained approximately 32.7% of the variance in CST 2011.

Research Question 4: Additional Variance Explained Beyond CST 2010

This section addresses the question: To what extend does fall performance on the DORF and Daze predict performance on the CST 2011, above and beyond the predictability of the previous year's CST score? A series of hierarchical analyses were run to address this question (see Table 10).

Hierarchical Regression. The first model examined the additional variance explained by DORF Fall alone. The second model examined additional variance explained by Daze Fall. The third model investigated the amount of additional variance explained when both DORF Fall and Daze Fall were entered into step 2.

In the first model, CST 2010 was entered in step 1 and was statistically significant, explaining 45.5% of the variance in CST 2011. After controlling for CST 2010 performance, there was a small, but significant change in R^2 [$F_{(2, 415)} = 197.87$, $p < .001$, $\Delta R^2 = .03$] when DORF Fall was entered in step 2. The final model was statistically significant and accounted for 48.8% of variance in CST 2011. Using the following regression model, CST 2011 = $B_0 + B_1$ CST 2010 + B_2 DORF Fall, the regression equation resulted in $Y' = 103.05 + .53X_{1i} + .36X_{2i}$.

In the second set of hierarchical models, CST 2010 was entered in step 1, followed by Daze Fall in step 2. Similar to the previous model, the addition of Daze Fall resulted in a small, but significant change in R^2 [$F_{(2, 356)} = 158.10$, $p < .001$, $\Delta R^2 = .02$]. The final model was statistically significant and explained 47% of the variance in CST 2011. The regression model, CST 2011 = $B_0 + B_1$ CST 2010 + B_2 Daze Fall, resulted in the following regression equation: $Y' = 100.10 + .59X_{1i} + 1.47X_{2i}$.

The last set of hierarchical models examined CST 2010 in step 1 and DORF Fall and Daze Fall in step 2. When DORF Fall and Daze Fall were entered simultaneously in step 2, there was a significant yet small change in R^2 [$F_{(2, 355)} = 119.12$, $p <.001$, $\Delta R^2 = .05$]. It is important to note when DORF Fall and Daze Fall were entered together in step 2, Daze Fall was non-significant ($p = .41$).

Research Question 5: Differences in Predictability for Varying EL Levels

Multiple regression analyses were conducted to answer the following question: To what extent does prediction bias play a role in the predictive validity of the DORF and the Daze, administered in the fall, for students of varying English language proficiency levels? Dummy variables were created to distinguish between students of different EL levels, resulting in 3 language proficiency groups. EA/A was chosen to serve as the reference group to allow for comparison against a group closest to fluent English language proficiency, or FEP status.

Multiple Regression. In the first analysis, DORF Fall and language proficiency were examined, yielding the following regression model: CST 2011 = B_0 + B_1 DORF Fall + B_2 B/EI + B_3 Int + B_4 B/EI*DORF Fall + B_5 Int*DORF Fall. DORF Fall was statistically significant in predicting CST 2011 (see Table 11). When DORF Fall was held constant, the addition of B/EI ($t = -6.10$, $p < .001$) and Int ($t = -2.04$, $p < .001$) were both significant, indicating there is a significant difference in CST 2011 performance due to English language proficiency level. The regression equation was as follows: $Y' = 303.57 + .53X_{1i} – 87.18X_{2i} – 26.67X_{3i} + .30X_{4i} – .09X_{5i}$. Students in the B/EI group are expected to score 87.18 points lower on the CST 2011 than students in the EA/A group, and students in the Int group are expected to score 26.67 points lower than students in the EA/A group. Both interaction effects were non-significant, suggesting that there is no difference in the predictive validity of the DORF Fall to CST 2011 for students in the B/EI and EA/A groups (*t* = 1.66, *p* = .10), nor for students in the Int and EA/A groups (*t* = -60, *p* = .55). The final model was significant and explained 51.6% of the variance in CST 2011.

In the second analysis, Daze Fall and language proficiency were examined, yielding the following regression model: CST 2011 = B_0 + B_1 Daze Fall + B_2 B/EI + B_3 Int + B_4 B/EI*Daze Fall + B_5 Int*Daze Fall. The model was significant and explained a total of 47.8% of the variance (see Table 12). It is important to note that once language proficiency was considered, Daze became non-significant (*t* = 1.57, *p* = .12). The difference in CST 2011 performance was significant for both the B/EI group (*t* = -9.44, *p* < .001) and the Int group (*t* = -5.5, *p* < .001). The following regression equation was yielded: $Y' = 340.70 + .1.28X_{1i} – 93.01X_{2i} – 48.11X_{3i} + 2.00X_{4i} + 1.05X_{5i}$. Students in the B/EI group tended to score 93.01 points lower on the CST 2011 than the reference group, EA/A students. Students in the Int group tended to score 48.11 points lower than students in the EA/A group. When examining the interaction effect between Daze Fall and B/EI (*t* = 1.5, *p* = .13), and Daze Fall and Int (*t* = 1.1, *p* = .28), results indicated that both interactions were non-significant. This suggests that the relationship between Daze performance and performance on the CST 2011 is the same for students in the B/EI and Int groups as for the students in the EA/A group.

Research Question 6:Predictive Accuracy for Different EL Groups

The diagnostic accuracy was calculated for each predictor variable to address the question of: What is the diagnostic accuracy of the DIBELS cut scores for DORF and Daze, in the fall, winter, and spring, for third grade students of differing language proficiency levels, as measured by sensitivity, specificity, positive predictive power, and negative predictive power? Current benchmark scores from DIBELS *Next* were used to define risk levels for DORF (Fall < 70, Winter < 86, Spring < 100) and Daze (Fall < 8, Winter < 11, Spring < 19). Based on state expectations, scores at the proficient or advanced level (≥ 350) on the CST were used to determine performance at or above expectations on the criterion measure.

As seen in Tables 13-14, for students in the B/EI group, the overall hit rate ranged between 69-85% on the DORF and 72-93% on the Daze. Sensitivity levels increased over time and were highest in the spring for both the DORF and the Daze. For the DORF, sensitivity levels increased from 69% in the fall, to 80% in the winter, to 85% in the spring. For the Daze, the sensitivity levels in the fall, winter, and spring were 72%, 86%, and 93%, respectively. It is important to note that none of the students in the B/EI group scored at or above expectations on the CST, resulting in scores of zero for false positives and zero for valid negatives for both the DORF and the Daze across all screening periods. In turn, the negative predictive power and specificity rates were 0%, while the positive predictive power was 100% during all screenings.

For students in the Intermediate group, the overall hit rate ranged between 56-65% on the DORF and 62-73% on the Daze (see Tables 15-16). Specificity rates for the

DORF in the fall and winter (81% and 71%) were higher than the sensitivity rates (54% and 64%) for the DORF. The sensitivity level was higher in the spring than the specificity level, 66% and 52%, respectively. Positive predictive power ranged between .93-.96 during the DORF screenings, indicating that there was a 93-96% probability that a student who was identified as being at risk on the DORF was truly at risk. Negative predictive power ranged between .14-.17 during the DORF screenings, indicating a 14-17% probability that students who were identified as not being at risk on the DORF were truly not at risk. Contrary to the DORF, sensitivity levels for the Daze were higher than specificity levels during each screening period. Sensitivity levels ranged between 63-74%, while specificity levels ranged between 38-63%. Positive predictive power ranged from 91-95%, and negative predictive power ranged from 11-19%.

For students in the EA/A group, sensitivity, specificity, positive predictive power, and negative predictive power were similar for both the DORF and the Daze (see Tables 17-18). The hit rate for the DORF ranged between 56-60%, increasing over time, while the hit rate for the Daze ranged between 55-61%. Sensitivity level for the DORF was low during all screenings, ranging between 26-47%. The sensitivity level of the Daze was slightly higher, but still low, ranging between 45-55%. Specificity ranged between 74-87% on the DORF and 65-70% on the Daze. Positive predictive power and negative predictive power of the DORF ranged between 65-67% and 53-58%, respectively. The Daze had a positive predictive rate between 57-63% and a negative predictive rate of 54-60%.

Chapter 5: Discussion

This study proposed to examine the effectiveness of DIBELS screening measures with third grade EL students of varying English language proficiency levels. Overall findings support previous research conducted with Spanish speaking ELs, indicating that the same measures used with native English speakers can also be used with this population of students (i.e., Baker & Good, 1995; Gersten et al., 2007). However, current DIBELS cut scores may need to be more closely examined with this group.

Results from this study extend the current literature on the use of fluency measures by focusing specifically on Spanish-speaking EL students of different levels of English proficiency. Multiple regression and hierarchical regression analyses were utilized to examine the predictive validity of the DORF and Daze to CST 2011 performance. As expected, there was an upward trend in scores as English language proficiency level increased for all measures. In comparison with the overall means of the entire sample, the B/EI students consistently scored lower, the Intermediate students scored around the overall mean, and the EA/A students performed above the overall mean.

Summary of Correlations

The correlations amongst the DORF assessments were significant in the fall, winter, and spring for all language proficiency groups, and the strength of association was similar across groups, ranging between .78-.90. These findings are consistent with previous studies that have examined DORF across different screening periods for third grade students (i.e., Roehrig et al., 2008; Shapiro et al., 2008). Shapiro et al. found correlations between DORF in the fall and winter to be .94, and Roehrig et al. found correlations ranging between .88 to .92 across screenings. The current findings with Spanish-speaking ELs are promising because correlations found between the DORF across screenings are very similar to studies that have been conducted with primarily native English speaking third grade students.

The relationship between the Daze during different screening periods was much lower, ranging between .26-.52. There has not been as much research conducted with the maze, and more specifically with the Daze, however, measures of comprehension have generally had lower correlations with each other than measures of oral reading fluency (i.e., Shapiro et al., 2008). When examining the difference in the strength of association for each language proficiency group, there was a trend indicating that the relationship amongst each Daze measure at different time points tended to be strongest for the B/EI group and weakest for the EA/A group. This could be attributed to the fact that students at the lower end of language proficiency tended to consistently score lower on the measures. Previous studies have shown that students who are poor readers tend to remain poor readers without consistent, targeted intervention (i.e., Juel, 1988). For students in the intermediate and advanced levels of language proficiency, performance was more variable from each screening period to the next, possibly resulting in lower overall correlations.

In general, DORF appeared to be more strongly correlated with all other measures than the Daze. One explanation for the weaker correlations with the Daze could be that the students were unfamiliar with the task. Many teachers reported that the students were unsure of what to do during the first screening, which may have negatively impacted the scores. In fact, many students scored low on the Daze in the fall but overall performance on this measure increased throughout the year, as did the general relationship between Daze performance and other measures, especially for students in the B/EI group.

The CELDT had a relatively low correlation with all measures. This was surprising because it was assumed that performance on the CELDT, which is supposed to represent EL level, would be highly correlated with all measures. It is possible that correlations were low based on the way overall CELDT score is calculated. The CELDT test takes into account listening, speaking, and writing, in addition to reading. Subsequently, EL level as determined by the CELDT, encompasses much more than just reading ability, whereas the fluency measures only assess reading ability.

Relationship of Fluency Measures and CST 2011 by EL Level

For the entire sample, the relationship between DORF and CST 2011 was statistically significant and strong across the screening periods. The relationship between scores on the DORF in the fall, winter, and spring with CST 2011 was consistently strongest for the B/EI group. Although still significant for students in the Int and EA/A groups, the DORF demonstrated a much weaker relationship with CST 2011 for these groups. This could be attributed to the fact that students in the B/EI group tended to score low from the screening measures to the criterion measure. Based on Stanovich (1986), this trend can be expected, as the high readers continue to grow and the low readers tend to stay the same.

The correlations across screening periods were similar within each language proficiency group. One would expect that the strength of association would be strongest between the spring fluency measures and the CST 2011 because the measures are administered so close in time. However, there was no distinction between strength of association across time points between CST 2011 and DORF. This is similar to results found in the study by Stage and Jacobson (2001), where the correlations between oral reading fluency given in the fall, winter, and spring with the Washington state test given in the spring were similar, ranging between .43-.44. Roehrig et al. (2008) also found similar correlations across screening periods, ranging from .66-.71.

Although there were moderate to strong correlations between DORF and CST 2011, previous research has often found much stronger correlations between oral reading fluency and state tests (i.e., Shapiro et al., 2008, Wood, 2006). It is important to consider that many studies were conducted with English only students and EL samples were usually small and often combined into one group. Wiley and Deno (2005) found a correlation of .61 between ORF and the Minnesota Comprehensive Assessment (MCA) for a mixed group of EL students in third grade. Their findings were similar to results from this study which found correlations up to .55 for the entire sample.

The relationship between Daze and CST 2011 generally increased across screening periods for both the entire sample and each EL group. This is consistent with previous studies that have shown that measures which are administered concurrently or within the smallest timeframe typically demonstrate the strongest relationship (i.e., Roehrig et al., 2008; Shapiro et al., 2006). Similar to the DORF, the relationship between

Daze and CST 2011 was stronger for students in the B/EI group than the Int and EA/A groups. That being said, Daze demonstrated only a small to moderate relationship with CST 2011 at most time points for all language proficiency groups. Results for third grade ELs in this study ranged between .39-.58, which is similar to findings from Wiley and Deno (2005), who found a correlation of .52 between the maze and MCA for EL students. Using standardized reading assessments as the outcome measures, rather than a state test, Ardoin et al. (2004) also found generally lower correlations between the maze and the outcome measures (.31-.51). Other studies conducted with mostly native English speakers have indicated stronger correlations for the maze, ranging between mid 60s-70s (Jenkins & Jewell, 1993). Although Jenkins and Jewell found that the maze was not very sensitive to individual differences among less skilled readers, findings in this study actually found a stronger relationship between the Daze and CST 2011 for students of lower reading ability (B/EI group).

The fact that the relationship between DORF and CST 2011, and Daze and CST 2011, was stronger for students in the B/EI group than for students of higher English language proficiency levels might be expected because students that demonstrate low reading skills, including comprehension, at the beginning of the year would be expected to demonstrate low comprehension skills at the end of the year. In contrast, students with higher reading ability are expected to show more growth over time (Stanovich, 1986).

Predictive Validity of Fall Fluency Measures

Findings from this study indicated that both the DORF and the Daze, administered in the fall, are significant predictors of performance of the CST 2011 in the spring. Overall, DORF appeared to be a better predictor and accounted for more variance than Daze. This is not surprising considering that a strong relationship between DORF and standardized reading achievement measures has been well established in previous research (i.e., Crawford et al., 2001; Reschly et al., 2009; Shapiro et al., 2008). The results are similar to results reported by Wiley and Deno (2005), who found that oral reading fluency was a better predictor than the maze on the state assessment for third grade EL students. Jenkins and Jewell (1993) also found that the maze was not as sensitive for less skilled readers, and appeared more sensitive to differences at upper grade levels.

When examining DORF and Daze in a hierarchical fashion, once DORF was accounted for, Daze did not add to the predictability of the CST 2011. In other words, results suggest that Daze may not be necessary when screening third grade Spanish speaking EL students once DORF is administered. When Daze was entered first, followed by DORF, DORF was able to explain a significant amount of additional variance, and Daze became non-significant. This is consistent with findings from Ardoin et al. (2004), where the addition of maze did not significantly improve R^2, nor explain significant unique variance beyond ORF. It is important to note that this study consisted predominately of native English speaking students. Wiley and Deno (2005) also found that the addition of maze for EL students was non-significant once oral reading fluency was accounted for.

These results are important because teachers often consider oral reading fluency to be less related to overall reading or comprehension ability (Fuchs et al., 1988).

Findings from this study support previous research indicating that oral reading fluency can be used to screen and identify students who may be at risk for low performance on state tests (i.e., Crawford et al., 2001; Reschly et al., 2009; Wood, 2006). One explanation for why the Daze was not as predictive as DORF is that past research has shown that oral reading fluency has a stronger relationship with overall reading than comprehension measures in the early grades. In fact, comprehension does not typically become a strong predictor of overall reading ability until fourth grade (Fuchs et al., 2001). Fuchs et al (2001) suggested that the relationship between oral reading fluency and reading comprehension is stronger in younger children and weakens as reading ability advances into more complex literary analyses. Jenkins and Jewell (1993) found the maze and oral reading fluency both had a strong relationship with reading achievement scores, but whereas the correlation between the maze and reading achievement did not show a decline as grade-level went up, the relationship between oral reading and achievement tests did decline as grade level increased. The current study only examined students in third grade, so it is possible that as the students move up in grade level, Daze may become a better predictor.

Predictability of Fall Fluency Measures Beyond Previous CST Scores

When examining if the DORF and Daze administered in the fall could contribute to predicting performance on CST 2011 in the spring for all Spanish ELs in third grade, results indicated that DORF was able to explain a small but significant amount of additional variance above and beyond the previous CST score. Daze was also able to explain significant additional variance but contributed less than the DORF. When both

DORF and Daze were entered together, the additional variance explained was significant, but the contribution of Daze was non-significant. Again, this suggests that the Daze may not be as useful in predicting outcomes for this population of students.

In the study by Wood (2006), when oral reading fluency was entered into the equation after the previous year's state test score, it explained an additional 9% of the variance in the current year's state test score. Together, previous year's CSAP score and oral reading fluency accounted for 62% of the variance. The author concluded that although oral reading fluency contributed only a small proportion of variance after prior year state test score, it still accounted for significant additional variance in the current year state test. The same findings were present in this study. Although the amount of additional variance accounted for with the addition of DORF in this study was slightly lower (3%), it still explained significant additional variance beyond that of the previous CST score. The fact that these quick 1-3 min fluency measures were able to explain significant variance above and beyond the previous year's CST scores is important because it will allow for further differentiation and more targeted instruction when allocating resources for interventions.

Difference in Predictive Validity of Fall Fluency Measures by EL Level

There was a significant difference in the level of performance on the CST 2011 for students of different language proficiency levels. As expected, students in the B/EI group performed much lower on the CST 2011 when compared to students in the EA/A group. Students in the Intermediate group also performed significantly lower than student in the EA/A group, but higher than students in the B/EI group.

More importantly, the interaction between DORF and EL level was non-significant, suggesting that there is no difference in the predictive ability of the DORF to CST 2011 based on English language proficiency. This finding is promising because it suggests that the DORF is able to predict performance on the CST 2011 at the same level for students of all English language proficiency levels. The same results were found with Daze. Again, the interaction effect was non-significant between Daze and each language proficiency group, indicating that there is no difference in the predictive ability of the Daze for students at the B/EI, Int, and EA/A levels.

The data suggests that language proficiency plays a large role in CST outcomes; however, it may not be a significant factor when examining the relationship between the fluency measures and CST performance. This is important because it suggests that DORF and Daze predict overall reading ability, as measured by the CST, similarly for third grade Spanish speaking ELs of all English language proficiency levels. As a result, findings from this study suggest that there is no evidence of predictive bias across English proficiency groups in this study. In other words, there does not appear to be a difference in the predictive validity of the DORF or the Daze in predicting CST 2011 outcomes for students at the beginning, intermediate, and advanced levels of English language proficiency. These results are similar to those of Betts et al. (2008), who did not find evidence of predictive bias between EL and native English speakers when examining fluency measures of phonemic awareness, alphabetic principle, and oral reading at the end of kindergarten as predictors of reading achievement at the end of second grade. The findings are also consistent with Roehrig et al. (2008), who through the use of logistic regression, found no significant interaction effects. They concluded that ORF was able to equally identify at risk readers regardless of demographic characteristics, including EL status.

In contrast to previous studies, Hosp et al. (2011) found variable predictive bias across grade and disaggregation categories, with most occurring during measurements at the beginning of the year. The authors suggested that the pattern of differential prediction could most likely be attributed to floor effects. Although many B/EI students in this study scored low on the Daze in the beginning of the year, the measure was still able to predict equally amongst groups.

Predictive Accuracy

Overall, the Daze had higher hit rates than the DORF during all screening periods and across all language proficiency groups. The hit rate generally increased over time, with the spring screening having the highest hit rate for both DORF and Daze. According to Rathvon (2004), screening measures often have high hit rates because they are able to predict which students will not become poor readers rather than predict which students will develop reading problems. This pattern was seen for students in the current study. The DORF and Daze were able to better predict students who were truly not at risk, but were not as accurate in predicting those who were truly at risk. Sensitivity levels proved to be much higher for students in the B/EI group and much lower for students in the EA/A group for both measures, indicating that the screening measures were able to predict truly at risk students better for groups with lower English language proficiency than for students with higher levels of English proficiency.

This is similar to findings by Buck and Torgesen (2003), who found that although minority students achieved benchmark scores for oral reading fluency, it was not as strong a predictor of success on the FCAT as it was for the white students. Conversely, the authors found that minority students with scores below benchmark were more likely to score below expectations on the FCAT than white students who scored below benchmark on ORF. The authors suggested that minority students typically have less developed reading skills, including vocabulary, which predict reading comprehension.

Discussion regarding the specificity of the DORF and Daze for B/EI students is limited, as all students in the B/EI group did not meet the expectation for the criterion measure. As a result, positive predictive power was 100% across all screenings for the DORF and Daze. Subsequently, specificity was zero, as was the negative predictive power across all screenings for both the DORF and Daze. The ability of the DORF in identifying truly at risk students was high during the winter and spring screening and met the 75-80% criteria established by Carran and Scott (1992), whereas the fall screening was not as accurate in identifying at risk students. For the B/EI students, the Daze was better than the DORF at correctly identifying at risk students across all screening periods.

The Intermediate group had more false negative rates than the B/EI and EA/A groups. This can be expected because students in the Int group tended to score more in the strategic need for support range on the DIBELS measures. As previous research has shown, it is generally more difficult to predict need for students who fall in the strategic range because this range is intended for students where an accurate prediction is not possible (Shapiro et al., 2008). Good et al. (2008) has stated that the DIBELS criteria are more accurate in classifying students at the ends of the reading distribution than students who fall in the middle. In general, the Daze had higher sensitivity and lower specificity levels across screening periods. Neither the DORF nor the Daze showed strong predictive accuracy, when using 75-80% as the standard (Carran & Scott, 1992).

For students in the Intermediate group, positive predictive power for the DORF and the Daze was relatively strong in terms of the proportion of valid positives in relation to all students identified as at risk. In other words, the number of false positives was minimal for this group. However, negative predictive power was extremely low due to a high number of false negatives. This is concerning because a high number of false negatives indicates that many students who are truly at risk have not been identified as at risk and therefore are not receiving the help they need.

Similar to students in the Intermediate group, sensitivity of the DORF was low across the fall, winter, and spring for students in the EA/A group due to a high number of false negatives. This is important because it suggests that maybe the current DORF criteria are too low for students with intermediate and advanced English proficiency levels in this study. Alternatively, the high number of false negatives associated with the DORF indicate that perhaps these EL students can read fluently but have difficulty with overall reading, or comprehension. Students with poor vocabulary and/or limited background knowledge, or second language learners often fall in this group (i.e., Bradley & Bryant, 1983; National Research Council, 1997). If this is the case, more research should be conducted to identify possible alternative measures.

Results from the current study are contrary to findings for the third grade EL group in the study by Hosp et al. (2011). Whereas the current study generally had higher specificity and lower sensitivity levels for both the DORF and the Daze, Hosp et al. found much higher sensitivity rates when examining ORF in relation to a state test for EL students in third grade. In their study, ORF demonstrated better specificity for non EL students, indicating that the measure was better at identifying which individuals in the non EL group would meet or exceed the criterion for proficiency. Hosp et al. suggested that when using ORF to screen third grade students, more false positive errors would occur for the EL group and more false negatives would occur for non ELs. Similar to Hosp et al., Shapiro et al. (2008) also found higher sensitivity levels than specificity levels on DORF for third grade students. In fact, both studies found sensitivity rates greater than 90%, while the current study found marginal sensitivity levels ranging from 26-85%.

Although the findings in this study are different from past studies, it is important to remember that there are many components that may contribute to differential prediction, including the predictor measures, the criterion measure, and the cut scores (Flaugher, 1978). First, cut scores established for DIBELS *Next,* as well as the updated passages, were used in the study. Shapiro et al. and Hosp et al. did not mention which version of DIBELS measures and cut scores were used in their study. Second, the criterion measure in this study differed from the criterion measure used in other studies, and is specific to the California state standards, which could be a contributing factor in the differences. Also, in the previous studies, ELs were examined as one group, regardless of ethnicity or English proficiency levels. The ELs in this study were disaggregated by EL level and limited to native Spanish speaking students.

Limitations

Ideally, a study examining the differences among students of differing English language proficiency levels would distinguish each group separately. Unfortunately, the number of students at the end levels (beginning and advanced) of English language proficiency were much lower than the number of students at the Intermediate level. Therefore, students in the end groups were combined in order to have a more meaningful comparison. Even after combining groups, the number of Intermediate students still outnumbered the other two groups, which may affect the ability to compare across groups due to differences in the consistency of scores and error for each group (Tabachnick & Fidell, 2007).

A second limitation is that no formal reliability checks were conducted. Although many teachers participated in multiple DIBELS assessment trainings, it was the first year that third grade teachers in this study implemented universal screening. Future studies should incorporate inter-rater data to ensure standardization of administration and accuracy of scoring.

Another possible limitation to this study is that many of the students in the sample may have been provided with supplemental reading interventions which may have influenced performance on subsequent screenings, as well as CST performance. However, it should be mentioned that providing interventions between screenings will likely be common practice as school districts move into consistent RtI implementation. In fact, the purpose of screening is to identify students who may be at risk so they can benefit from early intervention.

Another limitation was that the sample was restricted specifically to Spanish speaking EL students. The demographic makeup of the school district in this study is uncommon, as it is almost entirely comprised of one ethnicity/dominant language. It might be advantageous for future studies to have a comparison group of native English speakers, or different EL subgroups, in order to analyze whether the cut scores and predictive validity differ between these groups.

Potential Application for Research and Practice

Most research that has been conducted on oral reading fluency and maze assessments has focused on native English-speaking students. In many studies, the oral reading fluency and maze passages have been taken from different basal readers, instead of from a standardized, research-based system. In addition, of all the studies that have examined the predictive validity of the DIBELS measures to state-mandated assessments, none have compared them to the California state test.

This study contributes to the literature by examining the relationship and predictive validity specifically using DIBELS measures with third grade Spanish speaking EL students of varying English language proficiency levels. Results of the analyses indicate that Daze may not add valuable information beyond DORF during universal screenings. This is important because for several reasons. First, teachers often state that time dedicated to assessment interferes with instructional time (Reschly et al., 2009). If it is determined that a measure, in this case the Daze, does not provide meaningful data, assessment time could be minimized by eliminating this measure. Additionally, grouping of interventions are often based on screening results. If the Daze does not provide useful information as to later reading success, again it may not be necessary to administer this measure, nor would it be ideal to base intervention needs on Daze performance. Taking into consideration that comprehension is typically an area of concern for EL students, it is important to find a tool that can consistently identify EL students of differing proficiency levels who may require additional support in the area of reading.

The finding that the predictive validity of the DORF and Daze is not significantly different for students of varying English proficiency levels is promising because this suggests that the measures do not discriminate between students of the lowest to students of the highest language proficiency levels. In other words, as more schools move toward schoolwide screening, they can be confident that these measures can predict at-risk Spanish-speaking EL students equally, despite differing language classification levels.

Although results from this study suggest that there is no prediction bias for Spanish speaking students of different EL levels, findings from the predictive accuracy analysis indicated that current DIBELS *Next* benchmark scores may be too low for this group, resulting in a high false negative rate. Future research should be conducted to further examine current cut scores with Spanish EL students of differing language proficiency levels. In addition, future studies should investigate more optimal cut scores for students of different native languages as well as students with different language proficiency levels.

Results from this study support much of the research that has been conducted with oral reading fluency and maze measures. Although the Daze was not found to be a strong measure with the group of third grade Spanish ELs in this study, more research should be done to explore the use of fluency measures with ELs at different grade levels and of different ethnic/language backgrounds. As EL enrollment continues to increase, so does the need to identify valid measures for accurate and efficient identification of students at-risk for reading failure.

References

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

Adams, G., & Brown, S. (2007). The Six-Minute Solution: A Reading Fluency Program. Longmont, CO: Sopris West.

AERA, APA, & NCME. (1999). Testing individuals of diverse linguistic backgrounds. In AERA/APA/NCME (Ed.), *Standards for educational and psychological testing* (pp. 91-97). Washington, D.C.: American Educational Research Association.

Allinder, R. M., & Oats, R. G. (1997). Effects of acceptability on teachers' implementation of Curriculum-based Measurement and student achievement in mathematics computation. *Remedial and Special Education*, *18,* 113-120.

Ardoin, S. P., Witt, J. C., Suldo, S. M., Connell, J. E., Koenig, J. L., Resetar, J. L., Slider, N. J., & Williams, K. L. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probes when conducting universal screening. *School Psychology Review, 33,* 218-233.

Artiles, A. J., Rueda, R., Salazar, J., & Higareda, I. (2005). Within-group diversity in minority disproportionate representation English language learners in urban school districts. *Exceptional Children, 71,* 283-300.

Baker, S. K., & Good, R. H. (1995) Curriculum-based measurement of English reading with bilingual Hispanic students: a validation study with second-grade students. *School Psychology Review, 24,* 561-578.

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

Ballard & Tighe (2012). *Carousel of IDEAS, 4th Edition.* Retrieved February 19, 2012 from http://www.ballard-tighe.com/products/eld/carousel/.

Betts, J., Reschly, A., Pickart, M., Heistad, D., Sheran, C., & Marston, D. (2008). An examination of predictive bias for second grade reading outcomes from measures of early literacy skills in kindergarten with respect to English-language learners and ethnic subgroups. *School Psychology Quarterly, 23,* 553-570.

Bradley, L., & Bryant, P. E. (1983). Categorizing sounds and learning to read: A causal connection. *Nature, 303,* 419–421.

Buck, J., & Torgesen, J. (2003). The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test. (Technical Report) Tallahassee, FL: Florida Center for Reading Research.

California Department of Education (2010). *California English Language Development Test- CalEdFacts*. Retrieved August 6, 2010, from http://www.cde.ca.gov/ta/tg/el/cefceldt.asp

California Department of Education. (2011a). *California English Language Development Test- Technical Report 2009-2010 Edition*. Retrieved August 6, 2010, from http://www.cde.ca.gov/ta/tg/el/documents/techrpt0910.pdf#search=celdt%20technical%20manual&view=FitH&pagemode=none

California Department of Education (2011b). *California Standards Tests Technical Report.* Retrieved August 22, 2011 from http://www.cde.ca.gov/ta/tg/sr/documents/csttechrpt2010.pdf

California Department of Education (2011c). *Explaining and Using 2010-2011 Summary Results.* Retrieved December 26, 2011 from http://www.cde.ca.gov/ta/tg/el/documents/celdtsmryrslts2011.pdf.

California Department of Education (2008). *Standardized Testing and Reporting (STAR) Program.* Retrieved May 2, 2009 from http://www.cde.ca.gov/ta/tg/sr/documents/intrprslts08.pdf.

California Department of Education (2011d). *Post-Test Guide Technical Information for STAR District and Test Site Coordinators and Research Specialists.* Retrieved December 26, 2011 from http://www.startest.org/pdfs/STAR.post-test_guide.2011.pdf.

Carnine, D. (1997). Instructional design in mathematics for students with learning disabilities *Journal of Learning Disabilities, 30,* 130-131.

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

Christenson, S. L., Ysseldyke, J. E., & Thurlow, M. L. (1989). Critical instructional factors for students with mild handicaps: An integrative review. *Remedial and Special Education, 10,* 21-31.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cranney, A. G. (1972). The construction of two types of cloze reading tests for college students. *Journal of Reading Behavior, 5,* 60-64.

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

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

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37,* 184-192.

Deno, S. L., Marston, D., Deno, D. D., & Marston, D. (1988). Reading Progress Monitoring Passages. Minneapolis, MN: Children's Educational Services.

Deno, S. L., Maruyama, G., Espin, C., & Cohen, C. (1990). Educating students with mild disabilities in general education classrooms: Minnesota alternatives. *Exceptional Children, 57,* 150-161.

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49,* 36-45.

Deno, S. L., Mirkin, P., Chiang, B., & Lowry, L. (1980). *Relationships among simple measures of reading and performance on standardized achievement tests* (Res. Rep. No. 20). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities.

Dynamic Measurement Group (2011). *DIBELS Next Assessment Manual*. Retrieved February 13, 2011 from http://www.dibels.org/.

Espin, C., Deno, S. L., Maruyama, G., & Cohen, C. (1989). The Basic Academic Skills Samples (BASS): An instrument for the screening and identification of children at risk for failure in regular education classrooms. Paper presented at the annual meeting of the American Educational Research Association.

Flaugher, R. L. (1978). The many definitions of test bias. *American Psychologist, 33,* 671-679.

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

Fuchs, L. S., & Fuchs D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21,* 45-58.

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

Fuchs, D., & Fuchs, L. S. (2001). Responsiveness to intervention: A blueprint for practitioners, policymakers, and parents. *Teaching Exceptional Children, 38,* 57-61.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement, using a reading maze task. *Exceptional Children, 58,* 436-450.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal, 28,* 617-641.

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

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9,* 20-29.

Fuchs, D. F., Fuchs, L. S., Simmons, D. C., & Mathes, P. G. (2008). Peer Assisted Learning Strategies (PALS). Nashville, TN: Vanderbilt.

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

Gersten, R., Carnine, D., & Woodward, J. (1987). Direct instruction research: The third decade. *Remedial & Special Education, 8*(6), 48-56.

Geva, E. (2000). Issues in the assessment of reading disabilities in L2 children: Beliefs and research evidence. *Dyslexia, 6,* 13-28.

Goffreda, C. T., DiPerna, J. C., & Pedersen (2009). Preventative screening for early readers: Predictive validity of the dynamic indicators of basic early literacy skills (DIBELS). *Psychology in the Schools, 46,* 539-552.

Good, R. H., & Jefferson, G. (1998). Contemporary perspectives on Curriculum-Based Measurement validity. In M. R. Shinn (Ed.), *Advanced applications of Curriculum-Based Measurement* (pp. 61-88). New York: Guilford.

Good, R. H., & Kaminski, R. A. (2002). *Dynamic Benchmark Assessment: Assessment of Big ideas in beginning reading*. Eugene, OR: Institute for the development of educational achievement, University of Oregon, College of Education.

Good, R. H., Kaminski, R. A., Cummings, K., Dufour-Martel, C., Petersen, K., Powell-Smith, K., et al. (2011a). *DIBELS next assessment manual*. Dynamic Measurement Group, Inc.

Good, R. H., Kaminski, R. A., Dewey, E. N., Wallin, Powell-Smith, K. A, & Latimer, R. J. (2011b). *DIBELS next technical manual: Draft.* Dynamic Measurement Group, Inc.

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

Good, R. H., Simmons, D. C., & Smith, S. (1998). Effective academic interventions in the United States: Evaluating and enhancing the acquisition of early reading skills. *School Psychology Review, 27,* 45–56.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26,* 499-510.

Guthrie, J. T., Siefert, M., Burnham, N. A., & Caplan, R. I. (1974). The maze technique to assess, monitor reading comprehension. *The Reading Teacher, 28,* 161-168.

Hamilton, C. R., & Shinn, M. R. (2003). Characteristics of word callers: An investigation of the accuracy of teachers' judgments of reading comprehension and oral reading skills. *School Psychology Review, 32,* 228-240.

Harris, R. J. (1975). A *primer of multivariate statistics.* New York: Academic

Helms, J. E. (2006). Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *American Psychologist, 61,* 859-870.

Hintze, J. M., Ryan, A. L., & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Processing. *School Psychology Review, 32(4),* 541-556.

Hixson, M. D., & McGlinchey, M. T. (2004). The relationship between race, income, and oral reading fluency and performance on two reading comprehension measures. *Journal of Psychoeducational Assessment*, *22,* 351-364.

Hoover, H.D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1996). Iowa Test of Basic Skills, Form M. Itasca, IL: Riverside Publishing.

Hosp, J. L., Hosp, M. A., & Dole, J. K. (2011). Potential bias in predictive validity of universal screening measures across disaggregation subgroups. *School Psychology Review, 40,* 108-131.

Jenkins, J. R (2003). Candidate measures for screening at-risk students. Paper presented at the NRCLD responsiveness-to-intervention symposium, Kansas City, MO. Retrieved February 26, 2012, from http://www.nrcld.org/symposium2003/jenkins/index.html.

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: reading aloud and maze. *Exceptional Children, 59,* 421-432.

Jimenez, R. T., Garcia, G. E., Pearson, D. P. (1996). The reading strategies of bilingual Latina/o students who are successful English readers: Opportunities and obstacles. *Reading Research Quarterly, 31,* 90-112.

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice, 24,* 174-185.

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

Kame'enui, E., & Simmons, D. (2001). Introduction to this special issue: The DNA of reading fluency. *Scientific Studies of Reading, 5,* 203-210.

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

Klein, J. R. & Jimerson, S. R. (2005). Examining ethnic, gender, language, and socioeconomic bias in oral reading fluency scores among Caucasian and Hispanic students. *School Psychology Quarterly, 20,* 23-50.

Knofczynski, G. T., & Mundfrom, D. (2008). Sample sizes when using multiple linear regression for prediction. *Educational and Psychological Measurement, 68,* 431-442.

Langdon, H. (1989). Language disorder or difference? Assessing the language skills of Hispanic students. Exceptional Children, 56 (2), 160-167.

Leafstedt, J. M., Richards, C. R., & Gerber, M. M. (2004). Effectiveness of explicit Phonological awareness instruction for at-risk English learners. *Learning Disabilities Research and Practice, 19,* 252-261.

Lesaux, N. K., Geva, E., Koda, K., Siegel, L. S., & Shanahan, T. (2008). Development of literacy in second-language learners. In D. August & T. Shanahan (Eds.), *Developing reading and writing in second-language learners* (pp. 27-59). New York: Routledge.

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

Limbos, M., & Geva, E. (2001). Accuracy of teacher assessments of ESL children at-risk for reading disability. *Journal of Learning Disabilities*, *34*, 136-151.

MacGinitie, W. H., Kamons, J., Kowalski, R. L., MacGinitie, R. K., & McKay, T. (1978). Gates-MacGinitie Reading Tests (2nd ed). Chicago: Riverside.

Maleyko, G. & Gawlik, M. A. (2011). No child left behind: what we know and what we need to know. *Education, 131,* 600-624.

Marston, D. (1989). Curriculum-based measurement: What is it and why do it? In M. Shinn (Ed.), Curriculum-based measurement: Assessing special children (pp. 18-78). New York: Guilford Press.

Maxwell, S. E. (2000). Sample size and multiple regression analysis. Psychological Methods, 5, 434-458.

McGlinchey, M. T., & Hixon, M. D. (2001). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33*, 193-203.

National Assessment of Educational Progress; NAEP, (2011). *Achievement gaps. How Hispanic and white students in public schools perform in mathematics and reading on the national assessment of educational progress.* Retrieved June 24, 2011 from http://nces.ed.gov/nationsreportcard/pdf/studies/2011459.pdf

National Clearinghouse for English Language Acquisition; NCELA, (2011). *State Profiles and Reports*. Retrieved June 24, 2011 from http://www.ed-data.k12.ca.us/Navigation/fsTwoPanel.asp?bottom=%2Fprofile.asp%3Flevel%3D04%26reportNumber%3D16

National Clearinghouse for English Language Acquisition; NCELA, (2009). *The growing numbers of English learner students: 2008-2009 poster.* Retrieved June 24, 2011 from http://www.ncela.gwu.edu/files/uploads/9/growingLEP_0809.pdf

National Institute for Literacy (2006). *Responsiveness to Intervention.* Retrieved March 27, 2009 from http://www.nifl.gov/pipermail/learningdisabilities/2006/000396.html.

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

National Research Council. (1997). *Improving schooling for language-minority children.* Washington, DC: National Academies Press.

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

Parker, R., Hasbrouck, J. E., & Tindal, G. (1992). The maze as a classroom-based reading measure: Construction methods, reliability, and validity. *The Journal of Special Education, 26,* 195-218.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Fort Worth, TX: Harcourt Brace.

Prescott, G. A., Balow, I. H., Hogan, T.P., & Farr, R. C. (1984). Metropolitan Achievement Tests (MAT-6). San Antonio, TX: The Psychological Corporation.

Rathvon, N. (2004). *Early reading assessment: A handbook for practitioners*. New York: Guilford Press.

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

Riedel, B. W. (2007). The relation between DIBELS, reading comprehensions, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42,* 546-562.

Robb, L. (2002). The myth: learn to read/read to learn. *Scholastic Instructor, 111,* 23-25.

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

Schmidt, F. L. (1971). The relative efficiency of regression in simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement, 31,* 699-714.

Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment, 24,* 19-35.

Shapiro, E. S., Solari, E., & Petscher, Y. (2008). Use of a measure of reading comprehension to enhance prediction on the state high stakes assessment. *Learning and Individual Differences, 18,* 316-328.

Shaw, R. & Shaw, D. (2002). DIBELS Oral Reading Fluency-Based Indicators of Third Grade Reading Skills for Colorado State Assessment Program (CSAP). (Technical Report) Eugene, OR: University of Oregon.

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

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

Silver, Burdett, & Ginn (1991). *Word of reading.* Morristown, NJ: Author.

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

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360-407.

Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics* (2nd ed.). Cambridge, MA: Harper & Row.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Needham Heights, MA: Allyn & Bacon.

Thomas, W. P., & Collier, V. P. (1997). *School effectiveness for language minority students.* Washington, DC: National Clearinghouse for Bilingual Education.

Tindal, G., Marston, D., & Deno, S. L. (1983). *The reliability of direct and repeated measurement* (Research Rep. 109). Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.

U.S. Department of Education (2007). *Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001.* Retrieved January 21, 2009 from http://www.ed.gov/ policy/elsec/guid/saaprguidance.pdf

Vanderwood, M. L., Linklater, D., Healy, K. (2008). Predictive accuracy of nonsense word fluency for English language learners. *School Psychology Review, 37,* 1-13.

Vanderwood, M. L. & Nam, J. E. (2007). Response to intervention for English language learners: Current development and future directions. In S.R. Jimmerson, M.K. Burns, & A.M. VanDerHeyden (Eds.). *Handbook of response to intervention: The science and practice of assessment and intervention.* (pp. 408-417). NY: Springer.

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

Wetherill, G.B. (1986). *Regression analysis with applications*. Chapman and Hall, London.

Wiley, H. I., & Deno, S. L. (2005). Oral reading and maze measures as predictors of success for English learners on a state standards assessment. *Remedial and Special Education, 26*, 207-14.

Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading, 5,* 211-239.

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

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III Test of of Achievement. Ithasca, IL: Riverside Publishing.

Table 1. Predictive Accuracy Table

Outcome Measure (CST 2011)	Predictor Variables (DORF, Daze)		
	At-Risk	Not At-Risk	Indices
Below Expectations	Valid Positive (VP)	False Negative (FN)	Sensitivity VP/(VP + FN)
At or Above Expectations	False Positive (FP)	Valid Negative (VN)	Specificity VN/(VN + FP)
Total	Positive Predictive Value VP/(VP + FP)	Negative Predictive Value VN/(VN + FN)	Hit Rate (VP +VN)/ (VP + FN + VN + FP)

Note. CST = California Standards Test; Daze = DIBELS Daze; DORF = DIBELS Oral Reading Fluency; VP = Valid Positive; FN = False Negative; FP = False Positive; VN = Valid Negative.

Table 2. DIBELS *Next* Summary of Benchmark Goals for Third Grade

Measure	Intensive	Strategic	Core
DORF Fall	0-55	56-69	70 and above
DORF Winter	0-68	69-85	86 and above
DORF Spring	0-80	81-99	100 and above
Daze Fall	0-5	6-7	8 and above
Daze Winter	0-7	8-10	11 and above
Daze Spring	0-14	15-18	19 and above

Note. DORF = DIBELS Oral Reading Fluency; Daze = DIBELS Daze.

Table 3. Descriptive Statistics: Overall and Disaggregated by EL Level

	Overall	B/EI	Int	EA/A
Source	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
DORF Fall	69.81 (27.04)	52.09 (25.76)	69.46 (22.69)	89.92 (25.28)
DORF Winter	79.64 (27.12)	61.67 (26.83)	79.55 (22.82)	99.05 (24.72)
DORF Spring	90.41 (31.39)	70.64 (32.25)	91.28 (28.16)	109.34 (25.98)
Daze Fall	6.53 (4.27)	3.93 (3.53)	6.63 (4.05)	8.76 (4.12)
Daze Winter	9.48 (5.44)	6.20 (4.29)	9.36 (4.99)	12.90 (5.53)
Daze Spring	14.73 (6.69)	9.92 (6.63)	14.98 (5.83)	19.22 (5.46)
CST 2010	334.31 (46.33)	291.10 (34.65)	333.80 (33.96)	380.07 (40.59)
CST 2011	307.86 (43.56)	262.56 (33.42)	307.49 (31.46)	350.55 (36.53)

Note. Overall = All participants, B/EI = Beginning/Early Intermediate, Int = Intermediate, EA/A = Early Advanced/Advanced; DORF = DIBELS Oral Reading Fluency; Daze = DIBELS Daze; CST = California Standards Test English Language Arts; SD = Standard Deviation.

Table 4. Correlation Matrix for Entire Sample (N = 522)

Source	DORF Fall	DORF Winter	DORF Spring	Daze Fall	Daze Winter	Daze Spring	CELDT	CST 2010	CST 2011
DORF Fall	1.00								
DORF Winter	.89*	1.00							
DORF Spring	.84*	.89*	1.00						
Daze Fall	.60*	.61*	.58*	1.00					
Daze Winter	.54*	.57*	.53*	.48*	1.00				
Daze Spring	.60*	.66*	.65*	.50*	.52*	1.00			
CELDT	.51*	.51*	.48*	.43*	.43*	.51*	1.00		
CST 2010	.58*	.56*	.53*	.46*	.41*	.54*	.67*	1.00	
CST 2011	.54*	.55*	.53*	.39*	.41*	.58*	.69*	.67*	1.00

Note. * p < .01. DORF = DIBELS Oral Reading Fluency; Daze = DIBELS Daze; CELDT = California English Language Development Test; CST = California Standards Test English Language Arts.

Table 5. Correlation Matrix for B/EI Group (N = 122)

Source	DORF Fall	DORF Winter	DORF Spring	Daze Fall	Daze Winter	Daze Spring	CELDT	CST 2010	CST 2011
DORF Fall	1.00								
DORF Winter	.87**	1.00							
DORF Spring	.87**	.90**	1.00						
Daze Fall	.64**	.60**	.56**	1.00					
Daze Winter	.44**	.42**	.43**	.50**	1.00				
Daze Spring	.63**	.63**	.69**	.45**	.52**	1.00			
CELDT	.13	.19*	.28**	.09	.12	.27**	1.00		
CST 2010	.39**	.31**	.37**	.24*	.23*	.38**	.08	1.00	
CST 2011	.59**	.53**	.55**	.35**	.46**	.64**	.27**	.41**	1.00

Note. * *p* < .05; ** *p* < .01. B/EI = Beginning/Early Intermediate; DORF = DIBELS Oral Reading Fluency; Daze = DIBELS Daze; CELDT = California English Language Development Test; CST = California Standards Test English Language Arts.

Table 6. Correlation Matrix for Intermediate Group (N = 291)

Source	DORF Fall	DORF Winter	DORF Spring	Daze Fall	Daze Winter	Daze Spring	CELDT	CST 2010	CST 2011
DORF Fall	1.00								
DORF Winter	.86**	1.00							
DORF Spring	.78**	.85**	1.00						
Daze Fall	.56**	.58**	.53**	1.00					
Daze Winter	.45**	.47**	.44**	.43**	1.00				
Daze Spring	.45**	.56**	.53**	.43**	.33**	1.00			
CELDT	.28**	.28**	.25**	.34**	.19**	.27**	1.00		
CST 2010	.47**	.44**	.41**	.33**	.17**	.37**	.33**	1.00	
CST 2011	.31**	.37**	.38**	.30**	.13*	.41**	.48**	.46**	1.00

Note. * *p* < .05; ** *p* < .01. DORF = DIBELS Oral Reading Fluency; Daze = DIBELS Daze; CELDT = California English Language Development Test; CST = California Standards Test English Language Arts.

Table 7. Correlation Matrix for EA/A Group (N = 109)

Source	DORF Fall	DORF Winter	DORF Spring	Daze Fall	Daze Winter	Daze Spring	CELDT	CST 2010	CST 2011
DORF Fall	1.00								
DORF Winter	.85**	1.00							
DORF Spring	.84**	.86**	1.00						
Daze Fall	.33**	.39**	.35**	1.00					
Daze Winter	.41**	.51**	.39**	.26*	1.00				
Daze Spring	.44**	.51**	.51**	.28**	.51**	1.00			
CELDT	.37**	.42**	.42**	.20*	.32**	.32**	1.00		
CST 2010	.36**	.39**	.38**	.32**	.26**	.29**	.57**	1.00	
CST 2011	.36**	.39**	.36**	.15	.24*	.36**	.49**	.51**	1.00

Note. * *p* < .05; ** *p* < .01. EA/A = Early Advanced/Advanced; DORF = DIBELS Oral Reading Fluency; Daze = DIBELS Daze; CELDT = California English Language Development Test; CST = California Standards Test English Language Arts.

Table 8. Simple Regression Model Predicting CST 2011from DORF Fall (N = 452) and CST 2011 from Daze Fall (N = 388)

	R^2	B	B	*F*
DORF Fall	.29*	.54*	.91*	181.85*
Daze Fall	.15*	.39*	4.1*	69.85*

Note. * *p* < .001. CST = California Standards Test English Language Arts; DORF = DIBELS Oral Reading Fluency.

Table 9. Hierarchical Regression Analyses Predicting CST 2011 From Fall Fluency Measures (N = 388)

		R^2	ΔR^2	Final β	Final B	*F*
Step 1:	DORF Fall	.32*	.32*	.57*	.95*	182.27*
Step 2:	Daze Fall	.33	.01	.10	1.0	93.44*
Step 1:	Daze Fall	.15*	.15*	.39*	.4.1*	69.85*
Step 2:	DORF Fall	.33*	.17*	.51*	.86*	93.44*

Note. * *p* < .001. CST = California Standards Test English Language Arts; DORF = DIBELS Oral Reading Fluency; Daze = DIBELS Daze.

Table 10. Hierarchical Regression Analyses Predicting CST 2011 From CST 2010 and Fall Fluency Measures (N = 418)

		R^2	Δ R^2	Final β	Final B	*F*
Step 1:	CST 2010	.46**	.46**	.68**	.65**	347.86**
Step 2:	DORF Fall	.49*	.03**	.22**	.36**	197.87**
Step 1:	CST 2010	.45**	.45**	.67**	.65**	297.01**
Step 2:	Daze Fall	.47*	.02*	.14*	1.47*	158.10**
Step 1:	CST 2010	.45**	.45**	.67**	.65**	297.01**
Step 2:	DORF Fall			.24**	.40**	
	Daze Fall	.50**	.05**	.39	.40	119.12**

Note. * *p < .01;* ** *p* < .001. CST = California Standards Test English Language Arts; DORF = DIBELS Oral Reading Fluency; Daze = DIBELS Daze.

Table 11. Multiple Regression Model Predicting CST 2011 From DORF Fall and Language (N = 452)

	R^2	B	β	*F*
DORF Fall		.53**	.31**	
B/EI		-87.18**	-.80**	
Int		-26.67*	-.30*	
DORF Fall * B/EI		.30	.18	
DORF Fall * Int		-.09	-.08	
Model	.52**			94.99**

Note. * *p* < .05; ** *p* < .001. CST = California Standards Test English Language Arts; DORF = DIBELS Oral Reading Fluency; B/EI = Beginning/Early Intermediate; Int = Intermediate.

Table 12. Multiple Regression Model Predicting CST 2011 From Daze Fall and Language (N = 388)

	R^2	B	β	*F*
Daze Fall		1.28	.12	
B/EI		-93.01*	-.83*	
Int		-48.11*	-.54*	
Daze Fall * B/EI		2.00	.11	
Daze Fall * Int		1.05	.11	
Model	.48*			69.88*

Note. * *p* < .001. CST = California Standards Test English Language Arts; Daze = DIBELS Daze; B/EI = Beginning/Early Intermediate; Int = Intermediate.

Table 13. Predictive Accuracy of DORF to CST 2011 Performance for B/EI Group (N = 71)

Outcome Measure (CST 2011)	Predictor Variable (DORF)		
	Fall Predictor Variable		
	At-Risk DORF < 70	Not At-Risk DORF ≥ 70	Indices
Below Expectations *N* = 71	VP = 49	FN = 22	Sensitivity = .69
At or Above Expectations *N* = 0	FP = 0	VN = 0	Specificity = 0
Total	PPV = 1.0	NPV = 0	Hit Rate = .69
	Winter Predictor Variable		
	At-Risk DORF < 86	Not At-Risk DORF ≥ 86	Indices
Below Expectations *N* = 71	VP = 57	FN = 14	Sensitivity = .80
At or Above Expectations *N* = 0	FP = 0	VN = 0	Specificity = 0
Total	PPV = 1.0	NPV = 0	Hit Rate = .80
	Spring Predictor Variable		
	At-Risk DORF < 100	Not At-Risk DORF ≥ 100	Indices
Below Expectations *N* = 71	VP = 60	FN = 11	Sensitivity = .85
At or Above Expectations *N* = 0	FP = 0	VN = 0	Specificity = 0
Total	PPV = 1.0	NPV = 0	Hit Rate = .85

Note. B/EI = Beginner/Early Intermediate; PPV = Positive Predictive Value; NPV = Negative Predictive Value; VP = Valid Positive; FN = False Negative; FP = False Positive; VN = Valid Negative; DORF = Oral Reading Fluency; CST = California Standards Test.

Table 14. Predictive Accuracy of Daze to CST 2011 Performance for B/EI Group (N = 71)

Outcome Measure (CST 2011)	Predictor Variable (Daze)		
	Fall Predictor Variable		
	At-Risk Daze < 8	Not At-Risk Daze ≥ 8	Indices
Below Expectations *N* = 71	VP = 51	FN = 20	Sensitivity = .72
At or Above Expectations *N* = 0	FP = 0	VN = 0	Specificity = 0
Total	PPV = 1.0	NPV = 0	Hit Rate = .72
	Winter Predictor Variable		
	At-Risk Daze < 11	Not At-Risk Daze ≥ 11	Indices
Below Expectations *N* = 71	VP = 61	FN = 10	Sensitivity = .86
At or Above Expectations *N* = 0	FP = 0	VN = 0	Specificity = 0
Total	PPV = 1.0	NPV = 0	Hit Rate = .86
	Spring Predictor Variable		
	At-Risk Daze < 19	Not At-Risk Daze ≥ 19	Indices
Below Expectations *N* = 71	VP = 66	FN = 5	Sensitivity = .93
At or Above Expectations *N* = 0	FP = 0	VN = 0	Specificity = 0
Total	PPV = 1.0	NPV = 0	Hit Rate = .93

Note. B/EI = Beginner/Early Intermediate; PPV = Positive Predictive Value; NPV = Negative Predictive Value; VP = Valid Positive; FN = False Negative; FP = False Positive; VN = Valid Negative; Daze = DIBELS Daze; CST = California Standards Test.

Table 15. Predictive Accuracy of DORF to CST 2011 Performance for Int Group (N = 217)

Outcome Measure (CST 2011)	Predictor Variable (DORF)		
	Fall Predictor Variable		
	At-Risk DORF < 70	Not At-Risk DORF ≥ 70	Indices
Below Expectations *N* = 196	VP = 105	FN = 91	Sensitivity = .54
At or Above Expectations *N* = 21	FP = 4	VN = 17	Specificity = .81
Total	PPV = .96	NPV = .16	Hit Rate = .56
	Winter Predictor Variable		
	At-Risk DORF < 86	Not At-Risk DORF ≥ 86	Indices
Below Expectations *N* = 196	VP = 125	FN = 71	Sensitivity = .64
At or Above Expectations *N* = 21	FP = 6	VN = 15	Specificity = .71
Total	PPV = .95	NPV = .17	Hit Rate = .65
	Spring Predictor Variable		
	At-Risk DORF < 100	Not At-Risk DORF ≥ 100	Indices
Below Expectations *N* = 196	VP = 130	FN = 66	Sensitivity = .66
At or Above Expectations *N* = 21	FP = 10	VN = 11	Specificity = .52
Total	PPV = .93	NPV = .14	Hit Rate = .65

Note. Int = Intermediate; PPV = Positive Predictive Value; NPV = Negative Predictive Value; VP = Valid Positive; FN = False Negative; FP = False Positive; VN = Valid Negative; DORF = Oral Reading Fluency; CST = California Standards Test.

Table 16. Predictive Accuracy of Daze to CST 2011 Performance for Int Group (N = 217)

Outcome Measure (CST 2011)	Predictor Variable (Daze)		
	Fall Predictor Variable		
	At-Risk Daze < 8	Not At-Risk Daze ≥ 8	Indices
Below Expectations *N* = 196	VP = 123	FN = 73	Sensitivity = .63
At or Above Expectations *N* = 21	FP = 10	VN = 11	Specificity = .52
Total	PPV = .92	NPV = .13	Hit Rate = .62
	Winter Predictor Variable		
	At-Risk Daze < 11	Not At-Risk Daze ≥ 11	Indices
Below Expectations *N* = 196	VP = 134	FN = 62	Sensitivity = .68
At or Above Expectations *N* = 21	FP = 13	VN = 8	Specificity = .38
Total	PPV = .91	NPV = .11	Hit Rate = .65
	Spring Predictor Variable		
	At-Risk Daze < 19	Not At-Risk Daze ≥ 19	Indices
Below Expectations *N* = 196	VP = 146	FN = 52	Sensitivity = .74
At or Above Expectations *N* = 21	FP = 7	VN = 12	Specificity = .63
Total	PPV = .95	NPV = .19	Hit Rate = .73

Note. Int = Intermediate; PPV = Positive Predictive Value; NPV = Negative Predictive Value; VP = Valid Positive; FN = False Negative; FP = False Positive; VN = Valid Negative; Daze = DIBELS Daze; CST = California Standards Test.

Table 17. Predictive Accuracy of DORF to CST 2011 Performance for EA/A Group (N = 93)

Outcome Measure (CST 2011)	Predictor Variable (DORF)		
	Fall Predictor Variable		
	At-Risk DORF < 70	Not At-Risk DORF ≥ 70	Indices
Below Expectations *N* = 47	VP = 12	FN = 35	Sensitivity = .26
At or Above Expectations *N* = 46	FP = 6	VN = 40	Specificity = .87
Total	PPV = .67	NPV = .53	Hit Rate = .56
	Winter Predictor Variable		
	At-Risk DORF < 86	Not At-Risk DORF ≥ 86	Indices
Below Expectations *N* = 47	VP = 16	FN = 31	Sensitivity = .34
At or Above Expectations *N* = 46	FP = 8	VN = 38	Specificity = .83
Total	PPV = .67	NPV = .55	Hit Rate = .58
	Spring Predictor Variable		
	At-Risk DORF < 100	Not At-Risk DORF ≥ 100	Indices
Below Expectations *N* = 47	VP = 22	FN = 25	Sensitivity = .47
At or Above Expectations *N* = 46	FP = 12	VN = 34	Specificity = .74
Total	PPV = .65	NPV = .58	Hit Rate = .60

Note. EA/A = Early Advanced/Advanced; PPV = Positive Predictive Value; NPV = Negative Predictive Value; VP = Valid Positive; FN = False Negative; FP = False Positive; VN = Valid Negative; DORF = Oral Reading Fluency; CST = California Standards Test.

Table 18. Predictive Accuracy of Daze to CST 2011 Performance for EA/A Group (N = 93)

Outcome Measure (CST 2011)	Predictor Variable (Daze)		
	Fall Predictor Variable		
	At-Risk Daze < 8	Not At-Risk Daze ≥ 8	Indices
Below Expectations *N* = 47	VP = 22	FN = 25	Sensitivity = .47
At or Above Expectations *N* = 46	FP = 14	VN = 32	Specificity = .70
Total	PPV = .61	NPV = .56	Hit Rate = .58
	Winter Predictor Variable		
	At-Risk Daze < 11	Not At-Risk Daze ≥ 11	Indices
Below Expectations *N* = 47	VP = 21	FN = 26	Sensitivity = .45
At or Above Expectations *N* = 46	FP = 16	VN = 30	Specificity = .65
Total	PPV = .57	NPV = .54	Hit Rate = .55
	Spring Predictor Variable		
	At-Risk Daze < 19	Not At-Risk Daze ≥ 19	Indices
Below Expectations *N* = 47	VP = 26	FN = 21	Sensitivity = .55
At or Above Expectations *N* = 47	FP = 15	VN = 31	Specificity = .67
Total	PPV = .63	NPV = .60	Hit Rate = .61

Note. EA/A = Early Advanced/Advanced; VP = Valid Positive; FN = False Negative; FP = False Positive; VN = Valid Negative; PPV = Positive Predictive Value; NPV = Negative Predictive Value; Daze = DIBELS Daze; CST = California Standards Test.