UNIVERSITY OF CALIFORNIA
RIVERSIDE

Pseudo-Words vs. Real Words: Predicting Reading Outcomes for Culturally and Linguistically Diverse Students

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

Master of Arts

in

Education

by

Dennis Trévaughn Sisco-Taylor

September 2012

Thesis Committee:

Dr. Michael L. Vanderwood, Chairperson
Dr. H. Lee Swanson
Dr. Gregory Palardy

The Thesis of Dennis Trévaughn Sisco-Taylor is approved:

_____

_____

_____
Committee Chairperson

University of California, Riverside

ABSTRACT OF THE THESIS

Pseudo-words vs. Real Words: Predicting Reading Outcomes
for Culturally and Linguistically Diverse Students
By

Dennis Trévaughn Sisco-Taylor

Master of Arts, Graduate Program in Education
University of California, Riverside, September 2012
Dr. Michael L. Vanderwood, Chairperson

This study investigated the utility of two early literacy screening measures, Nonsense Word Fluency (NWF) and Word Identification Fluency (WIF), with a sample of culturally and linguistically diverse students. Included within the overall sample of 196 first grade students were 106 Spanish-speaking ELLs at varying levels of English proficiency. Screening measures were administered in the winter of first grade, and used to predict oral reading fluency (ORF) at the end of the school year. Results indicated that WIF accounted for substantial variance in ORF above and beyond that accounted for by NWF. Receiver operating characteristic curve analyses revealed that WIF was superior to NWF in terms of classification accuracy. Differences in predicting ORF for ELLs at low, moderate, and high levels of English proficiency were observed on the NWF measure. Study outcomes provide support for WIF as an early literacy screener within early prevention and intervention frameworks.

Table of Contents

Introduction……1

Method……9

Results……13

Discussion……...19

References……...26

Figures and Tables……...33

Pseudo-words vs. Real Words: Predicting Reading Outcomes for Culturally and Linguistically Diverse Students

Research outcomes spanning across the past three decades have suggested that early intervention is necessary for children who struggle to acquire early literacy skills (Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; Fuchs et al., 2001; Juel, 1988; Stanovich, 1986). Students that show deficits in phonemic awareness (PA) and phonics in their first years of school are likely to develop into poor readers and remain so throughout their educational tenure (Juel, 1988; Stanovich, 1986). Unfortunately, this is a fate that has become all too familiar for children that are ethnic minorities and come from low socioeconomic backgrounds. Significant gaps in reading achievement continue to exist between White students and other racial subgroups (Hemphill & Vanneman, 2011; Vanneman, Hamilton, Baldwin Anderson, & Rahman, 2009). Reports from the National Assessment of Educational Progress (NAEP) have shown that only 14% of students from high poverty schools perform at or above a proficient level in reading by 4th grade (Vanneman et al., 2009). Since African American and Hispanic students make up the majority of enrollments in high poverty schools, these numbers paint an incredibly stark picture for these two subgroups. In addition to the aforementioned ethnic subgroups, high poverty schools also tend to have large populations of English language learners (ELLs). Approximately 25% of the students in high poverty schools are considered ELL (Vanneman et al., 2009). The NAEP report (2009) also disclosed that about three-quarters of ELLs were native Spanish speakers. This particular subgroup is deserving of a

great deal of attention since a 29-point gap in reading achievement exists between them and their native English-speaking (NES) Hispanic counterparts (Hemphill & Vanneman, 2011).

Multi-tiered preventative frameworks such as response-to-intervention (RTI) have been frequently cited as system approaches that can be used to address the previously outlined problems (Deno et al., 2009; Fuchs, Mock, Morgan, & Young, 2003; Marston, Muyskens, Lau, & Canter, 2003). They can be used to identify struggling students, and provide them with appropriate instructional supports (Fuchs, Mock, Morgan, & Young, 2003; Marston, Muyskens, Lau, & Canter, 2003). Perhaps the most critical component of these systems is universal screening, as it represents the first step in a preventative approach. When screening is universal, all students have been administered a reliable and valid measure that has been shown to predict outcomes in a given area of interest (Deno et al., 2009). Students that perform below a pre-determined benchmark are generally considered at-risk for failing high stake assessments (Deno et al., 2009). Quality screening measures share two key attributes: criterion validity and classification accuracy (see Jenkins, Hudson, & Johnson, 2007 for review). One such example, are measures of oral reading fluency (ORF). These measures have a great deal of utility since they can serve as a general outcome measure (GOM) in reading and track reading development across time (Fuchs, Fuchs, & Compton, 2004). However, oral reading fluency measures become less useful when the population of concern consists of students that do not have well developed reading skills (Bain & Garlock, 1992; Catts, Petscher, Schatsneider, Bridges, & Mendoza, 2009). As demonstrated by Catts and colleagues (2009), the ORF

measure has significant floor effects when administered to students in the first grade. Floor effects can make it difficult to gauge a student's current level of reading ability and more importantly, correctly classify them as at-risk or not at-risk for failing high stake assessments. Accordingly, in order to optimize screening in the first grade or earlier, measures that target lower order skills (e.g., PA, phonics) are more ideal. Nonsense word fluency (NWF) and word identification fluency (WIF) are two fluency-based measures of phonics skill that have received a lot of attention in the screening literature. The former task involves reading pseudo-words in isolation while the latter involves reading high frequency sight words in isolation. The paper will proceed with a brief literature review on both approaches to measuring phonics.

Measuring Phonics through Pseudo and Real Word Reading Tasks

The utility of pseudo-word reading tasks in distinguishing between skilled and less skilled readers was documented by Perfetti and Hogaboam (1975). In this investigation, skilled readers had shorter vocalization latencies in comparison to less skilled readers when decoding real words and pseudo-words. However, the biggest differences between skilled and less skilled readers were observed with pseudo-words. Word identification tasks also have a long and established history within the CBM literature. Deno, Mirkin, and Chiang (1982) identified reading words in isolation as a behavior that could be used to measure progress in reading and predict future reading outcomes. Ehri's work (e.g., Ehri, 2005) suggests that word identification tasks have this capability because they can reflect a child's mastery of the alphabetic principle. Under this theoretical framework, readers that are able to recognize whole words instantly have

reached a consolidated alphabetic phase where unitization is the common approach for reading unknown words (Harn, Stoolmiller, & Chard, 2008).

Regardless of the approach used to measure phonics skills, getting an accurate snapshot of phonics acquisition in students' first years in school is important since decoding is the strongest predictor of reading ability in younger children (Gough, Hoover, & Peterson, 1996). As expressed in Perfetti and Hogaboam (1975), readers requiring considerable processing resources to decode single words will have fewer resources available for higher order processes, such as reading comprehension. For these reasons, sight word and pseudo-word reading measures such as WIF and NWF are ideal for screening purposes. Accordingly, the following sections will provide an in-depth review of the literature examining the efficacy of both measures as early literacy screeners.

WIF as an Early Literacy Screener

Fuchs, Fuchs, and Compton (2004) compared WIF to NWF in predicting various reading outcomes at the end of first grade. In this study, fall WIF scores were stronger predictors of fluency and comprehension in the spring than fall NWF scores. Building on the work of Fuchs and colleagues, Compton, Fuchs, Fuchs, and Bryant (2006) investigated the utility of WIF as an academic screener, and progress monitoring device. When WIF was added to a first grade screening battery that included sound matching, rapid digit naming, and oral vocabulary, sensitivity and specificity rates approaching .85 and .81 respectively were produced. However, adding WIF to the aforementioned screening battery did not significantly increase the area under the curve (AUC) in the

receiver operating characteristic (ROC) curve analysis. Thus, WIF level alone was not able to significantly improve classification accuracy of the screening battery. In a more recent investigation by Clemens, Shapiro, and Thoemmes (2011), WIF was compared to phonemic segmentation fluency (PSF), letter naming fluency (LNF), and NWF in predicting reading outcomes at the end of first grade. WIF emerged as the strongest predictor of reading outcomes at the end of first grade. WIF was also superior to the other screening measures in terms of classification accuracy, producing the largest AUC of the four measures.

Although WIF emerged as a robust measure of first grade reading outcomes in this study, the general conclusion was that it did not produce large enough rates of specificity that would warrant its use as an independent screener within an RTI framework. However, combining WIF with other screening measures did result in modest improvements in classification accuracy. The literature synthesis conducted by Wayman and colleagues (2007) revealed that word identification measures were strong predictors of passage fluency four months later. Together, these findings provide support for the use of these measures with younger English proficient students that have not yet become fluent readers. However, there is a need for more research to be conducted on these measures with CLD students, especially ELLs. After all, practices that are to be used in school settings that host CLD learners should first be validated with research samples that are representative of these student populations (AERA, APA, & NCME, 1999; Klingner & Edwards, 2006; Vanderwood & Nam, 2008). To date, this has not been the case with

word identification measures as a review of the literature did not reveal any studies addressing screening and classification accuracy specifically with CLD subgroups.

The potential for these measures to be differentially effective in predicting reading outcomes for different subgroups must be taken seriously since there is at least some evidence that suggests that other types of oral reading measures (e.g., oral reading fluency) could have this tendency (Hixson & McGlinchey, 2004; Klein & Jimerson, 2005). Also, given the strong correlation between sight word reading fluency and oral reading fluency of connected text, it seems at least plausible that these two tasks may produce similar errors in predicting reading outcomes.

NWF as an Early Literacy Screener

The psychometric properties of NWF were examined in a recent literature review conducted by Goffreda and DiPerna (2010). Test-retest reliability coefficients for NWF proved to be adequate for screening purposes, and NWF demonstrated moderate to high levels of concurrent validity with other reading measures across studies. Predictive validity levels also ranged from moderate to high. The highest of these correlations were with the DIBELS oral reading fluency (DORF) measure. Notwithstanding, evidence relating to the sensitivity and specificity of the NWF measure was mixed.

Vanderwood, Linklater, and Healy (2008) examined the predictive utility of the NWF measure with an ELL sample. First grade NWF scores were better predictors of later reading achievement than state achievement tests (e.g., SAT-9). NWF scores from the end of first grade had moderate to strong relationships with third grade literacy outcomes. Harn, Stoolmiller, and Chard (2008) reported a restriction of range, where fall

NWF scores lacked predictive utility for students whose initial score was greater than 49 correct letter sounds per minute. Fien, Baker, Smolkowski, Smith, Kame'enui, and Beck (2008) examined the criterion validity of NWF with first grade ELLs and native English-speakers (NESs). Fall NWF scores were predictive of spring ORF and SAT-10 scores for both ELLs and NESs. Winter NWF scores were able to account for between 74- 76% of the variances respectively in ORF for ELLs and NESs. This was not a statistically significant difference.

Jenkins, Hudson, and Johnson's (2007) review of early literacy measures revealed that NWF has shown poor sensitivity, especially when predicting the bottom quartile on achievement tests. Johnson, Jenkins, Petscher, and Catts (2009) added that NWF showed poor specificity when unsatisfactory reading was defined as below the 40th percentile. NWF demonstrated only 50% specificity when sensitivity was set at 90%. These rates were even lower for ELLs, which prompted the suggestion that different cut scores are likely needed for these groups (Johnson et al., 2009). Johnson et al. (2009) also found that NWF produced a classification accuracy rate of 75%. However, this was only a 4% increase over the base rate, the overall classification accuracy that could have been obtained by skipping the screening process altogether. In their review of the DIBELS measures, Catts and colleagues (2009) reported that NWF showed significant floor effects and poor predictability when it was first administered with students in winter of kindergarten, as recommended by DIBELS. Fortunately, the floor effects were reduced by the fourth administration of the measure (i.e., the winter benchmark of 1st grade). Not

surprisingly, improvement in predictability at the lower quartiles of the predictor variables mirrored reductions in the floor effects.

Research Objectives

The purpose of this study is to add to the existing literature on early screening measures by evaluating the efficacy of NWF and WIF with a CLD student population that includes Spanish-speaking ELLs. This investigation should also provide more information pertaining to the classification accuracy of these two screening measures, and thus the feasibility of using them within early prevention frameworks. The investigation is guided by the following research questions: (1a) To what extent do sight word and pseudo-word reading measures account for additional variance in ORF beyond each other in the total sample? (1b) To what extent do sight word and pseudo-word reading measures account for additional variance in ORF beyond each other in the ELL subsample? (2) Do measures of sight word fluency and pseudo-word reading produce rates of classification accuracy that would be acceptable within a direct route RTI model when R-CBM is used as the outcome variable? (3) Do pseudo-word and sight word reading measures predict oral reading fluency differently for ELL and NES students? (4) For ELLs, are there significant within group differences between students with low, moderate, and high levels of English proficiency (EP) in predicting ORF with either screening measure (NWF or WIF)? (4a) Which measure, NWF or WIF, predicts ORF best for ELLs?

Method

Participants

Participants were (N = 195) students from ten first grade classrooms in two urban elementary schools in southern California that implement a RTI framework. All first grade students were screened as part of the schools' universal screening efforts. Male students accounted for approximately 51% of the sample (n = 99). In terms of ethnicity, approximately 82% of the students were Hispanic (n = 159) while approximately 18% were African American (n = 35). One student was of Samoan origin. Spanish speaking ELLs (n = 106) made up approximately 54% of the sample. All other students (n = 89) were considered native English speakers (NES).

All students that were classified as ELLs were administered the California English Language Development Test (CELDT; California Department of Education, 2004). The CELDT yields scores on a scale of 1-5 (1 = beginning, 2 = early intermediate, 3 = intermediate, 4 = early advanced, 5 = advanced) that can be interpreted in an ordinal fashion and describe the students' level of English proficiency. The subsample of ELL students had the following distribution in terms of English proficiency: approximately 7% fell in the beginning range (n = 7), 19% in the early intermediate range (n = 20), 46% in the intermediate range (n= 49), 26% in the early advanced range (n = 28), and 2% in the advanced range (n = 2). For the purposes of this study, CELDT groups were merged to create three levels of English proficiency (low, moderate, and high). The beginning and early intermediate groups were combined to form the low group, and the early advanced and advanced groups were merged to form the high group. Thus, for the

purposes of this investigation, students with CELDT levels of 1-2 were considered to have low English proficiency, while level 3 was considered moderate, and levels 4-5 were considered high.

Free or reduced-price lunch (FRPL) status was used as an index for determining socioeconomic status (SES) in this study. The National Assessment of Educational Progress (2009) defined high poverty schools as those with 76-100% of students qualifying for FRPL. In this study, 100% of the students from both schools qualified for FRPL indicating that these were high poverty schools serving low SES populations.

Measures

AIMSweb Nonsense Word Fluency (NWF). The AIMSweb NWF measure was created based on the testing practices operationalized by DIBELS (see Good & Kaminski, 1998; Shinn & Shinn, 2002). During this task, students are presented with an 8.5" x 11" sheet of paper with randomly ordered VC and CVC nonsense words. Students are asked to say all of the sounds in the pseudo-word or read the whole word, and encouraged to say any sounds that they know in the word. The total number of correct letter sounds are calculated after one minute. NWF shows moderate concurrent validity (.59) with the Woodcock Johnson Psycho-Educational Battery-Revised readiness cluster in February of first grade, and strong predictive validity (.82) with end of year ORF when administered in January of first grade (Good & Kaminski, 2002). It has also demonstrated strong one month alternate-form reliability (.83) in the winter of first grade (Good & Kaminski, 2002).

Word Identification Fluency (WIF). During the WIF task, students were presented with a list of 80 high frequency sight words on a single sheet of paper. The words were sampled from the Dolch preprimer, primer, and first-grade-level lists (see Fuchs, Fuchs, & Compton, 2004). Students were asked to read as many of the sight words as possible within one minute. Students were prompted to move on to the next word on the list when hesitating for more than four seconds. The total number of words correct per minute were calculated by taking the total number of correctly produced words, dividing that by the actual time (in seconds), then multiplying the quotient by 60 In Fuchs, Fuchs, and Compton (2004), concurrent validity between WIF and the word identification subtest of the Woodcock Reading Master Test (WRMT) was .77. Alternate test-form reliability was reported to be .97 from two consecutive weeks of administration.

Reading Curriculum-based Measurement (R-CBM) probes. The AIMSweb R-CBM probes are measures of oral reading fluency (ORF) where students read connected text for a one-minute period (Shinn & Shinn, 2002).. Bain and Garlock (1992) reported a criterion validity coefficient of .62 with their 1st grade sample. The R-CBM score is the total number of words read correctly in one minute (WCPM). In this study, three R-CBM probes were administered to each student, and the median words correct per minute from the three probes was used as the outcome variable. R-CBM has demonstrated alternate form reliability of .89 for 1st grade probes (Shinn & Shinn, 2002).

California English Language Development Test (CELDT). The CELDT is a test of English language proficiency that is administered to all students (K-12) in the state of California whose primary language is not English, and students that were previously

classified as English learners and have not been reclassified as fluent-English proficient (California Department of Education, 2002). The CELDT determines the level of language proficiency and assesses student progress in acquiring listening, speaking, reading, and writing skills (California Department of Education, 2002). Students are assessed at the beginning of each school year. According to the California Department of Education (2002) overall performance for students in grades K-1 is calculated on the following scale: 45% listening, 45% speaking, 5% reading, and 5% writing. Internal consistency estimates of reliability for the reading and writing subtests range from .85-.91 (California Department of Education, 2002). The CELDT classifies students at the following performance levels: beginning, early intermediate, intermediate, early advanced, and advanced (California Department of Education, 2002).

Procedures

Data were collected in two waves (winter and spring benchmark periods) by the author, and two school employees, all of whom received training on administering the academic screening measures from certified AIMSweb trainers. To ensure that measures were being administered with fidelity, inter-rater agreement was calculated on approximately 10% of the probes. The mean inter-rater agreement was calculated at 95% for all of the screening measures. NWF screening data were collected in January shortly after students returned from winter break. WIF data were collected in February, approximately four weeks after the NWF data. In order to control for time differences between the administration of the two measures across students, data were collected from the respective classrooms in the same order in each wave of data collection. The R-CBM

measures were administered in May, at the end of the school year. Each student was administered three grade-level R-CBM probes, and the median correct words per minute from the probes were used as the outcome variable in the data analyses. The order of presentation of the R-CBM probes was randomly counterbalanced across students to control for order effects.

Results

In order to evaluate the efficacy of WIF and NWF as early literacy screeners with CLD student populations, data were collected from 195 CLD first grade students. Data analyses were conducted to address the proposed research questions. Findings for each research question will be discussed in the following sections.

Descriptive Analysis

Before addressing any of the proposed research questions, a descriptives procedure was conducted to evaluate the fit of the data for parametric statistical analyses. The mean, standard deviation, skewness, and kurtosis calculations for the total sample on all of the CBM measures are displayed in Table 1. As reflected in the table, the sample distribution on the NWF measure appears to be leptokurtic and have a slight positive skew ($\gamma_2 = 2.28$, $\gamma_1 = 1.37$) while the distributions for WIF and R-CBM appear to be approximately normal. Mean and standard deviation calculations for the ELL sample were sorted by the level of language proficiency and are displayed in Table 2.

Since multiple regression procedures were used to address the majority of the research questions, it was necessary to check the statistical assumptions for linear regression. In checking these statistical assumptions, slight violations of normality were

observed as evidenced by the skewness and kurtosis statistics, and a normal probability plot of regression standardized residuals. As discussed in Pedhazur (1997), violations of normality can impact the estimation of coefficients.

Research Question 1

To address the first research question regarding the extent to which sight word and pseudo-word reading accounted for significant variance in ORF above and beyond each other in the total sample and ELL subsample, the following regression models were specified:

(1) $Y_{\text{R-CBM}} = \beta_0 + \text{NWF} + \text{WIF} + \text{NWF*WIF} + e$.

(2) $Y_{\text{R-CBM}} = \beta_0 + \text{WIF} + \text{NWF} + \text{NWF*WIF} + e$.

In these regression models, $Y_{\text{R-CBM}}$ represents ORF measured by R-CBM, WIF and NWF represent the beta coefficients for sight-word and pseudo-word reading, WIF*NWF represents the interaction effect between sight-word and pseudo-word reading, and *e* represents the residual term. The interaction between NWF and WIF was not significant, and was thus excluded from subsequent models. The absence of a significant interaction indicates that the main effects for NWF and WIF can be interpreted independently (Pedhazur, 1997). As discussed in Pedhazur (1997), the inclusion of irrelevant variables decreases the degrees of freedom (*df*), and can increase the standard error for relevant variables included within the model.

A hierarchical regression procedure was conducted where the individual predictors were entered in the model one at a time (e.g., NWF, followed by WIF, followed by the interaction term). In the first model, where NWF was entered in the

model first, the adjusted R^2 value increased from approximately 42% to 86% when the WIF predictor was added, an R^2 change of 44%. NWF and WIF were significant predictors of ORF when included in the prediction model together, demonstrating that WIF accounted for substantial variance in ORF above and beyond that of NWF. In the second model, when WIF was entered in the model first followed by NWF, the adjusted R^2 value increased from 85% to 86%, an R^2 change of approximately 1% when NWF was added. Both predictors remained significant when NWF was added to the regression model.

To address the second part of the first research question, the extent to which sight word and pseudo-word reading accounted for significant variance in ORF above and beyond each other in the ELL subsample, the same regression models were tested with NES students excluded from the analyses. In the first model, where NWF was entered first, the adjusted R^2 increased from approximately 54% to 90% when WIF was added to the model, an increase of approximately 36%. Both NWF and WIF were significant while included in the model together, reflecting that each measure contributed significant variance to the prediction of ORF when accounting for the contributions of the other. When WIF was entered first in the second model, the adjusted R^2 value increased from approximately 86% to 90%, an R^2 change of approximately 4%. Both NWF and WIF were significant predictors when included in the model together.

Research Question 2

In an effort to address the second research question, pertaining to the classification accuracies of NWF and WIF when trying to predict reading outcomes for

CLD students, two separate receiver operating curve (ROC) analyses were conducted, one for each screening measure. In conducting the ROC analyses, the first task was to select an objective criterion for unsatisfactory reading performance in the spring of 1st grade. Unsatisfactory reading performance was defined as a median score from three 1st grade *AIMSweb* R-CBM passages that fell below the 25th percentile, which corresponds with 40 words read correctly per minute in the spring of 1st grade. Thus, students reading less than 40 words correctly per minute would be identified as at-risk for future reading difficulties. For each of the measures, cut scores yielding approximately 90% sensitivity were identified. These cut scores, along with corresponding specificity rates, and area under the curve (AUC) estimates are displayed in Table 3.

The NWF measure yielded approximately 53% specificity when sensitivity was set at 91%, and accounted for approximately 82% of the area under the curve (AUC). Thus, when using a cut score that yielded approximately 91% sensitivity, NWF would correctly classify 53% of the students that had satisfactory reading performance as not at-risk. The WIF measure was able to yield 91% specificity when sensitivity was set at 90%, accounting for 98% of the AUC.

Research Question 3

In addressing the third research question regarding whether pseudo-word reading or sight word reading tasks predicted ORF differently for ELLs and NESs, two regression models were tested. The first regression model, $Y_{\text{R-CBM}} = \beta_0 + \text{NWF} + \text{ELL} + \text{NWF*ELL} + e$, (where ELL represented a dummy-coded variable for language status, and NWF*ELL represented the interaction between pseudo-word reading and language

status) was able to account for significant variance in ORF, $F(3, 173) = 44.85$, $p < .01$, adjusted $R^2 = .43$. The interaction between pseudo-word reading and language status was not significant, however pseudo-word reading and language status were both significant predictors within the model. The beta coefficient for ELLs was significantly lower than that of non-ELLs, indicating that there were differences in ORF between ELLs and NESs after controlling for pseudo-word reading level. The second model, $Y_{\text{R-CBM}} = \beta_0 + \text{WIF} + \text{ELL} + \text{WIF*ELL} + e$, accounted for even more variance in ORF, $F(3, 178) = 391.22$, $p < .001$, adjusted $R^2 = .87$. No interaction between sight word reading and language status was observed, however. Additionally, the intercepts for ELLs and non-ELLs were not significantly different. The WIF measure was the only significant predictor in the model, indicating that the sight word reading task predicted similarly for both groups, and that there were no mean differences in ORF after controlling for sight word reading skill.

Research Question 4

To address the fourth research question regarding whether either screening measure predicted ORF differently for ELLs at low, moderate, and high levels of English proficiency (EP), two regression models were tested; one for each of the screening measures:

(1) $Y_{\text{R-CBM}} = \beta_0 + \text{NWF} + \text{lowEP} + \text{moderateEP} + \text{NWF*lowEP} + \text{NWF*moderateEP} + e$.

(2) $Y_{\text{R-CBM}} = \beta_0 + \text{WIF} + \text{lowEP} + \text{moderateEP} + \text{WIF*lowEP} + \text{WIF*moderateEP} + e$.

In these models, lowEP and moderateEP represented dummy-coded variables for low and moderate levels of EP. The high EP group was used as the reference group. In the first model, the NWF*moderateEP interaction term was significant. In addition, both dummy

variables were significant, indicating that when pseudo-word reading scores were accounted for in the model, the ORF performances of ELLs at low and moderate levels of EP were significantly different than those of ELLs with a high level of EP. The intercept coefficients for the low and moderate groups are negative and significant at an alpha level of .01, indicating that the ORF performance of both of these groups was significantly less than that of the high EP group when holding pseudo-word reading skill constant. The significant interaction between NWF and the moderate group indicated that NWF predicted ORF differently for ELLs at a moderate level of EP in comparison to ELLs at a high level of EP. The positive slope coefficient ($\beta = .29$, $p < .05$) reflected that NWF was a stronger predictor of ORF for the moderate EP group in comparison to the high EP group. The interaction term for the low EP group was not significant, indicating that NWF predicted ORF similarly for the low and high groups. This linear relationship is displayed in Figure 1.

In the second model, sight word reading skill was the only significant predictor of ORF in the model. The interaction terms were not significant, and the dummy variables for language proficiency level did not account for significant variance in ORF. This indicates that the WIF measure predicted similarly for ELLs at each level of EP, and that after accounting for sight word reading skill there were no significant differences in ORF for ELLs at the three levels of EP.

In addressing the second part of the research question, regarding which screening measure predicted ORF best for ELLs, Fisher's *r*-to-*z* transformation (Rosenthal & Rosnow, 1991) was used. As reflected in Table 4, WIF was the stronger predictor of ORF

for ELLs. The correlation coefficient for WIF (r = .93) was significantly greater (z = -4.88, p < .001) than the correlation coefficient for NWF (r = .74). This trend is consistent with what was observed for the total sample. WIF showed greater predictive validity with R-CBM across both subgroups. For NES students, an even greater disparity between correlation coefficients was observed, r = .58 and r = .93 for NWF and WIF respectively.

Discussion

The intended purpose of this investigation was to add to the existing literature on two early literacy screening measures, NWF and WIF. Of particular interest was their utility in predicting reading outcomes for students from CLD backgrounds. Consistent with past research (e.g., Clemens, Shapiro, & Thoemmes, 2011; Fuchs, Fuchs, & Compton, 2004), WIF was a stronger predictor of reading outcomes at the end of first grade in comparison to NWF. While both NWF and WIF were significant predictors of ORF, the data reflected that WIF contributed a substantial amount of the variance in ORF even after accounting for the contribution of NWF. An examination of the hierarchical regression models revealed a change in adjusted R^2 of approximately .44 when WIF was added to NWF. On the other hand, very minimal gains were observed when NWF was added to WIF. For the total sample, only a one percent increase in the adjusted R^2 was observed. While this one percent increase was statistically significant, accounting for 86% of the variance in ORF in comparison to 85% may carry very little practical significance, especially when the time to administer each of these measures to a student population is taken into consideration.

In examining the classification accuracies of the two measures, data from the ROC analyses reflected that WIF yielded results that would be more acceptable within an early prevention framework. Using the criteria for evaluating AUC discussed in Compton et al. (2010), where AUC greater than .90 is considered excellent; .80 to .90, good; .70 to .80 fair; and below .70, poor, the AUC produced by WIF would be considered excellent while NWF's, good. WIF also demonstrated 91% specificity when holding sensitivity at approximately 90%. While there is no definitive criterion for sensitivity, there has been at least some agreement within the screening literature (e.g., Jenkins et al., 2007; Johnson et al., 2009) that sensitivity should be held at .90 to limit the number of false negatives screening instruments will produce. Limiting the number of false negatives accrued during the screening process is paramount to an early prevention framework as it ensures that the majority of truly at-risk students will in fact be identified as at-risk and receive the appropriate accommodations (Jenkins et al., 2007). While there is a lack of consensus over what constitutes adequate sensitivity, there is even less agreement among researchers as to what constitutes an acceptable threshold for specificity. However, recommendations from Compton et al. (2010) suggest that specificity rates at .80 would be acceptable. The specificity rates produced by NWF when holding sensitivity at 91% would not be deemed acceptable by this criterion. On the other hand, WIF displayed specificity rates over .90 when holding sensitivity rates at .90. This outcome provides support for the use of WIF within an RTI framework.

NWF vs. WIF: Predicting ORF for ELLs and NESs

While there have been a couple of investigations examining NWF with ELLs (e.g., Fien et al., 2008; Vanderwood et al., 2008), a review of the literature did not produce any studies that have examined the WIF measure with Spanish-speaking ELLs. In this study, both measures predicted ORF similarly for ELLs and NESs. This is consistent with the findings of Fien and colleagues (2008) who reported no differences on NWF in predicting ORF outcomes for the respective groups. WIF emerged as the stronger predictor of ORF for both subgroups.

Mean differences in ORF between ELLs and NESs were observed even after controlling for NWF level. Thus, when eliminating the variance associated with pseudo-word reading skill (a measure of phonics acquisition) there were still differences in ORF performance. Specifically, ELLs had an average that was 10 words correct per minute (WCPM) less than NESs when the variance associated with NWF was accounted for. These mean differences in ORF between ELLs and NESs were not observed when accounting for sight word reading performance, however. One possible explanation for this finding is that sight word reading tasks may be measuring a number of different early literacy skills in addition to phonics, as suggested in Aaron, Joshi, Ayotollah, Ellsbury, Henderson, and Lindsey (1999). Also, as expressed in Ehri (2005), the ability to rapidly decode sight words could represent the level of one's mastery of the alphabetic principle.

In exploring the differences between ELLs at different levels of English Proficiency (EP) for the two screening measures, some notable findings emerged. First, there were no differences between ELLs at the different levels of English proficiency

(EP) in predicting ORF with the WIF measure. In addition, no mean differences between the respective groups were observed on ORF after accounting for the variance associated with sight word reading. Differences between ELLs at the different levels of EP emerged when predicting ORF with the NWF measure, however. The data suggest that NWF was a stronger predictor for ELLs at the moderate level of EP in comparison to the high group. No significant differences in predicting ORF between the low and high groups was observed, however. This was likely due to measurement error in estimating the slope for the low EP group. While the slope coefficient for the low EP group was similar to that of the moderate group, the standard error estimate for the low EP group was larger, which may explain why the interaction was not statistically significant.

Similar to the trend observed between ELLs and NESs, there were mean differences in ORF at the respective levels of EP after accounting for variance associated with pseudo-word reading. The high EP group had an average ORF score that was 16 units greater than the moderate group, and 27 units greater than the low group. From a practical standpoint, this means that even after accounting for differences in decoding skill as measured by NWF, students that had a high level of EP were still far superior readers than other ELL students. Together, these findings suggest that sight word reading tasks may account for variance in reading connected text that cannot be explained purely by phonics skills (pseudo-word decoding). If this is in fact the case, it would support Ehri's position on reading where children reach a consolidated phase of reading and no longer use decoding as their primary mechanism for reading text.

Implications for Practice

Consistent with findings from past investigations (e.g., Clemens, Shapiro, & Thoemmes, 2011; Compton, Fuchs, Fuchs, & Bryant, 2006; Fuchs, Fuchs, & Compton, 2004), the WIF measure was an extremely powerful predictor of reading outcomes at the end of first grade. Findings from the current investigation suggest that the WIF measure's predictive power is not limited to students that represent the ethnic majority. Instead, preliminary evidence presented in this study suggests that the WIF measure can be just as useful with CLD student populations. Together, these findings present a strong case for including the WIF measure in first grade screening batteries. Findings from this study support past research (e.g., Clemens, Shapiro, & Thoemmes, 2011; Fuchs, Fuchs, & Compton, 2004) suggesting that WIF is a stronger predictor of reading outcomes at the end of first grade in comparison to NWF. However, NWF continues to be a very widely used measure. An estimated 735, 000 students were administered the AIMSweb NWF measure in 2010 (AIMSweb, 2010). This figure does not include the number of students that were administered the DIBELS NWF measure, which is likely much greater due to its availability. While NWF presents both predictive and instructional utility as a screener, consumers should be aware that it may not predict similarly for all students.

Following the recommendations of Vanderwood and Nam (2008), ELLs were not viewed as a homogeneous group in this study. Instead, Spanish-speaking ELLs were examined at different levels of English proficiency. While they all share a common obstacle in learning a new language, there are many within group differences. ELLs come from different cultures, hold different values, speak different languages, and range on a

continuum of English proficiency. Differences in reading achievement were observed in this study, even within this subset ELL students. These findings suggest that the level of English proficiency should be taken into consideration when projecting reading outcomes for ELLs. Thus, when working with ELLs, educators should look beyond their language status and consider other variables, such as their language proficiency.

Limitations

In examining the outcomes of this study and its' potential contributions to the literature, a number of limitations should be considered. First, R-CBM was the one and only outcome variable used in the study. While R-CBM, DIBELS oral reading fluency (DORF; see Good et al., 2011), and other measures of ORF are recognized as global outcome measures of reading, results may not generalize to high stake assessments as well as those yielded from traditional norm-referenced achievement tests of reading (e.g., WRMT, SAT-10). Also, using a multivariate approach with multiple outcome measures of reading would likely result in more accurate representations of students' overall reading achievement. A second limitation of the study was the sample size. A larger sample of ELLs would have likely resulted in more accurate estimates in the regression procedures used within the study. The power analysis reflected that an *n* of 138 would have been needed to detect a moderate effect in a regression model with five predictors. A third limitation of the investigation was the regression of spring reading outcomes on winter benchmark predictors. Most screening investigations predict spring reading outcomes with fall performance on early literacy measures. The levels of specificity observed herein may have been closer to those observed in other investigations if fall

benchmark data were used. Also, from a practical standpoint, one would not want to wait until the winter benchmark period to begin intervening with struggling readers. However, past research (e.g., Catts et al., 2009) has suggested that the NWF measure (and other DIBELS measures) presents significant floor effects up until the fourth administration, which would fall at the winter benchmark period of first grade. Thus, one could make the case that it is easier to discriminate between students at different risk levels in the winter of first grade with NWF. This was the rationale used within the current investigation.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Testing individuals of diverse linguistic backgrounds. In, *Standards for educational and psychological testing* (pp. 91-97). Washington, DC: Author.

Aaron, P. G., Joshi, R. M., Ayotollah, M., Ellsbury, A., Henderson, J., & Lindsey, K. (1999). Decoding and sight-word naming: Are they independent components of word recognition skill? *Reading and Writing, 11*, 89-127.

AIMSweb. (2010). *AIMSweb growth table*. Retrieved November 2, 2011 from http://www.aimsweb.com/

Bain, S. K., & Garlock, J. W. (1992). Cross-validation of criterion-related validity for CBM reading passages. *Diagnostique, 17*, 202-208.

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*(4), 561-578.

Blachman, B. A., Ball, E. W., Black, R. S., & Tangel, D. M. (1994). Kindergarten teachers develop phoneme awareness in low-income, inner-city classrooms. *Reading and Writing: An International Journal, 6*, 1-18.

California Department of Education. (2002). *Technical report for the California English Language Development Test (CELDT)*. Retrieved September10, 2011, from www.cde.ca.gov

Catts, H. W., Petscher, Y., Schatschneider, C., Bridges, M. S., & Mendoza, K. (2009). Floor effects associated with universal screening and their impact on the early identification of reading disabilities. *Journal of Learning Disabilities, 42*(2), 163-176.

Clemens, N. H., Shapiro, E. S., & Thoemmes, F. (2011). Improving the efficacy of first grade reading screening: An investigation of word identification fluency with other early literacy indicators. *School Psychology Quarterly, 26*(3), 231-244.

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., et al. (2010). Selecting at-risk first grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology, 102*(2), 327-340.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology, 98*(2), 394-409.

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*(1), 36-45.

Deno, S. L., Reschly, A. L., Lembke, E. S., Magnusson, D., Callender, S. A., Windram, H., & Stachel, N. (2009). Developing a school-wide progress-monitoring system. *Psychology in the Schools, 46*(1), 44-55.

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading, 9*(2), 167-188.

Fien, H., Baker, S. K., Smolkowski, K., Smith, J. L. M., Kame'enui, E. J., & Beck, C. T. (2008).Using nonsense word fluency to predict reading proficiency through second grade for English language learners and native English speakers. *School Psychology Review,37*(3), 391-408.

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90*(1), 37-55.

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2004). Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children, 71*(1), 7-21.

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

Goffreda, C. T., & DiPerna, J. C. (2010). An empirical review of psychometric evidence for the Dynamic Indicators of Basic Early Literacy Skills. *School Psychology Review, 39*(3), 463-483.

Good, R. H., Baker, S. K., & Peyton, J. A. (2009). Making sense of nonsense word fluency: Determining adequate progress in early first-grade reading. *Reading & Writing Quarterly, 25*(1), 33-56.

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

Good, R. H., Kaminski, R. A., Cummings, K. D., Dufour-Martel, C., Petersen, K., Powell-Smith, K., et al. (2011). *DIBELS Next*. Eugene, OR: DMG. Available at http://dibels.org/.

Graves, A. W., Plasencia-Peinado, J., Deno, S. L., & Johnson, J. R. (2005). Formatively evaluating the reading progress of first-grade English learners in multiple-language classrooms. *Remedial and Special Education, 26*(4), 215-225.

Harn, B. A., Stoolmiller, M., & Chard, D. J. (2008). Measuring the dimensions of alphabetic principle on the reading development of first graders: The role of automaticity and unitization. *Journal of Learning Disabilities, 41*(2), 143-157.

Hemphill, F. C., & Vanneman, A. (2011). *Achievement gaps: How Hispanic and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress* (NCES 2011-459). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

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.

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*(4), 582-600.

Johnson, E. S., Jenkins, J. R., & Petscher, Y. (2010). Improving the accuracy of a direct route screening process. *Assessment for Effective Intervention, 35*(3), 131-140.

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*(4), 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.

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*(1), 23-50.

Klingner, J. K., & Edwards, P. A. (2006). Cultural considerations with response to intervention models. *Reading Research Quarterly, 41*(1), 108-117.

Linan-Thompson, S., Vaughn, S., Prater, K., & Cirino, P. T. (2006). The response to intervention of English language learners at risk for reading problems. *Journal of Learning Disabilities, 39*(5), 390-398.

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

Marston, D., Muyskens, P., Lau, M., & Canter, A. (2003). Problem-solving model for decision making with high-incidence disabilities: The Minneapolis experience. *Learning Disabilities Research & Practice, 18*(3), 187-200.

National Assessment of IDEA Overview (NCEE 2011-4026). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U. S. Department of Education.

National Institute of Child Health and Human Development (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: US. Government Printing Office.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Holt, Rinehart, & Winston.

Perfetti, C. A., & Hogaboam, T. (1975). Relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology, 67*(4), 461-469.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Shapiro, E. S., & Clemens, N. H. (2009). A conceptual model for evaluating system effects of response to intervention. *Assessment for Effective Intervention, 35*(1), 3-16.

Shinn, M. R., & Shinn, M. M. (2002). *AIMSweb training workbook*. Eden Prairie, MN: Edformation, Inc.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquistion of literacy. *Reading Research Quarterly, 21*(4), 360-407.

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34*(1), 33-58.

Vanderwood, M. L., Linklater, D., & Healy, K. (2008). Predictive accuracy of nonsense word fluency for English language learners. *School Psychology Review, 37*(1), 5-17.

Vanderwood, M. L., & Nam, J. (2008). Best practices in assessing and improving English language learners' literacy performance. In A. Thomas & J. Grimes (Eds). *Best Practices in School Psychology V*. Bethesda, MD: National Association of School Psychologists.

Vanneman, A., Hamilton, L., Baldwin Anderson, J., & Rahman, T. (2009). *Achievement gaps: How Black and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress,* (NCES 2009-455). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Vaughn, S., Mathes, P. G., Linan-Thompson, S., & Francis, D. J. (2005). Teaching English language learners at risk for reading disabilities to read: Putting research into practice. *Learning Disabilities Research & Practice, 20*(1), 58-67.

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

Table 1

Descriptive Statistics for CBM Measures (Total Sample)

Source	*M*	*SD*	Skewness (γ_1)	Kurtosis (γ_2)
NWF (N=189)	57.15	34.03	1.37	2.28
WIF (N=194)	35.54	23.81	0.41	-0.90
R-CBM (N=180)	53.92	35.05	0.59	-0.38

Table 2

Descriptive Statistics for English Language Learners by Level of Language Proficiency

		R-CBM	NWF	WIF
Low EP		N = 25	N = 27	N = 27
	M	27.92	39.67	17.78
	SD	19.93	17.94	15.22
Moderate EP				
		N = 46	N = 49	N = 49
	M	47.30	53.12	37.57
	SD	27.26	26.32	21.94
High EP		N = 27	N = 30	N = 30
	M	78.70	85.07	52.77
	SD	28.36	45.16	17.68
Total		N = 98	N = 106	N = 106
	M	51.01	58.74	36.83
	SD	31.88	35.58	23.02

Table 3

Classification Indices for Spring of 1st Grade

Source	AUC	Sensitivity	Specificity	Cut Scores
NWF	.82	91%	53%	39
WIF	.98	90%	91%	31

Note. Sensitivity was set at 91% for NWF since the ROC analysis did not produce a cut score that was consistent with 90% sensitivity.

Table 4
Predictive Correlations of NWF and WIF with ORF by Language Status

Measure	*n*	*r*	*z*-score
NWF_{ELL}	98	.74	-4.88**
WIF_{ELL}	98	.93	
NWF_{NES}	76	.58	-6.12**
WIF_{NES}	81	.93	

Note. Correlations between WIF, NWF, and R-CBM are from Winter of 1st grade to Spring of 1st grade. ** = $p \leq .01$, * = $p \leq .05$.

Figure 1

Predicting Spring R-CBM with Winter NWF at Different Levels of English Proficiency