UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

An Examination of the Predictive Validity of Early Literacy Measures for Spanish-Speaking English Language Learners in First Grade

Permalink

https://escholarship.org/uc/item/7wk2m8z7

Author

Hatch, Abigail Ann

Publication Date

2018

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
RIVERSIDE

An Examination of the Predictive Validity of Early Literacy Measures for Spanish-Speaking English Language Learners in First Grade

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

Doctor of Philosophy

in

Education

by

Abigail A. Hatch

March 2018

Dissertation Committee:
Dr. Rollanda O’Connor, Chairperson
Dr. H. Lee Swanson
Dr. Austin Johnson

Copyright by
Abigail A. Hatch
2018

The Dissertation of Abigail A. Hatch is approved:

____

____

____

Committee Chairperson

University of California, Riverside

Acknowledgments

This dissertation was the final step in completing my doctorate, and I would like to take this opportunity to thank some of the people who made a profound impact on me during graduate school. To my initial advisor, Dr. Mike Vanderwood, thank you for admitting me to the PhD program and introducing me to the response-to-intervention approach, both in theory and in practice. My research interests and focus as a school psychologist have been shaped by what I learned from you.

I would also like to thank the staff and students in Compton Unified School District. I appreciate the lessons you taught me, the many opportunities you gave me to put into practice what I was studying, and the patience you showed an enthusiastic but inexperienced consultant.

To my students in EDUC 267, 172, and 110, thank you for your eagerness to learn and your heart for teaching. You helped me stay connected to the many reasons I was working on my doctorate and reminded me how much it mattered.

To my dissertation committee members, Dr. Austin Johnson and Dr. Lee Swanson, thank you for generously giving your time and feedback to improve this paper. Most of all, thank you to my dissertation chairperson, Dr. Rollanda O'Connor, for guiding me through revising my proposal and finally completing my dissertation. Without your patience and your research expertise, this dissertation would never have been written. I am in your debt.

Dedication

Among the many people I have to thank, two people stand out. This dissertation is dedicated to my father, Ronald Hatch. Your achievements as a scientist have always inspired me, and though I would not have chosen the circumstances we have been dealing with over the last several years, I will always be grateful for the opportunity to get to know you better as an adult. Thank you for the sacrifices you have made to help me finish even while we both cared for Mom.

This dissertation is also dedicated to my friend, Yesenia Cesena. Thank you for being there through the ups and downs. You have been both a shoulder to cry on and a cheerleader to celebrate with. As I finish school and you go back, I am excited to see what lies ahead for both of us.

ABSTRACT OF THE DISSERTATION

An Examination of the Predictive Validity of Early Literacy Measures for
Spanish-Speaking English Language Learners in First Grade

by

Abigail A. Hatch

Doctor of Philosophy, Graduate Program in Education
University of California, Riverside, March 2018
Dr. Rollanda O'Connor, Chairperson

This study examined the predictive validity and classification accuracy of four early literacy screeners with Spanish-speaking English language learners (Ss ELLs) at varying levels of English language proficiency. First grade Ss ELLs (N = 209) were screened in the fall and winter using Phoneme Segmentation Fluency (PSF), Developmental Spelling Test (DST), Nonsense Word Fluency (NWF), and Word Identification Fluency (WIF). The criterion measure administered in the spring was Oral Reading Fluency (ORF). Overall, the WIF screener was the strongest predictor of end-of-first grade ORF scores. At each assessment period, it was the most strongly correlated with ORF3 for the total sample and for students aggregated by language proficiency groups. Results of hierarchical regression models, with PSF, DST and DST^2, NWF, and WIF entered in order, found that each addition resulted in a significant change in R^2 with respect to spring ORF scores. However, once all screeners were entered, WIF was the only significant predictor of ORF3 in the winter regression model, and one of two (with NWF) significant predictors in the fall. Language proficiency did not explain additional variance in ORF3 when added to the regression model after the literacy screeners. When it comes to classification accuracy, WIF had the largest area under the curve (AUC) of the fall and winter measures and the highest specificity (when sensitivity was selected as close to .90 as possible). Though it was not as strong as WIF's, DST's classification accuracy was much better than that of PSF. Findings indicate that the same cut scores can be used for ELLs at each level as for native English speakers (NESs). Results from this study contribute to the research suggesting that schools should use both WIF and a spelling measure, like DST, to assess first graders.

Table of Contents

List of Tables

Chapter 1: Introduction

In 2011, the White House Initiative on Educational Excellence for Hispanics and the U.S. Department of Education released the report, *Winning the Future: Improving Education for the Latino Community.* The authors pointed out that Hispanics are the "largest and fastest growing minority group in the U.S. yet have the lowest education attainment levels" (p. 2). Data from the U.S. Census Bureau give further details. Their report on educational attainment stated that, in 2015, 88 percent of adults aged 25 and older were at least high school graduates, while among Hispanic adults, only 67 percent were (Ryan & Bauman, 2016). Similarly, 33 percent of adults had a bachelor's degree or more, while only 16 percent of Hispanic adults did. While there are many factors that contribute to this gap in educational attainment, research indicates that one key factor is earlier difficulties with reading comprehension (August & Shanahan, 2006; Snow, Burns, & Griffin, 1998).

Biennial assessments are conducted nationwide as part of the National Assessment of Educational Progress (NAEP). Students in grades 4 and 8 are assessed in both reading and mathematics, and data from these assessments are published and referred to as the "Nation's Report Card." Over the course of data collection from 1992 to 2009, there has been a documented achievement gap between Hispanic and White students, despite growth in reading scores for both groups (National Center for Education Statistics; NCES, 2011). In the most recent report on reading achievement (NCES, 2011), the authors stated that, overall, eight percent of fourth graders scored in the Advanced range, 26 percent in Proficient, and 33 percent in the Basic range. This means that 33 percent (one-third) of fourth graders were below the Basic range and were not even able to demonstrate the “partial mastery of prerequisite knowledge and skills” (p. 6) that is required for the Basic achievement level. Among Hispanic students in fourth grade, 49 percent scored below the Basic range.

Research indicates that lack of English language proficiency is a factor that impacts student achievement (Halle, Hair, Wandner, McNamara, & Chien, 2012; Snow et al., 1998). While not all Hispanic students are English language learners (ELLs), a considerable proportion are. Data collected by the U.S. Department of Education indicated that 37 percent of Hispanic students in grade 4 and 21 percent in grade 8 were ELLs (NCES, 2011). As a part of the No Child Left Behind Act (NCLB, 2002), schools were held accountable for the performance of all students, including specific subgroups. ELLs are one subgroup whose performance has been below average. Statistics from NAEP indicated that 70% of fourth grade ELLs scored Below Basic compared to 30% of native English speakers (NESs; NCES, 2011). In fact, while the achievement gap described previously that exists between Hispanic students and White students is a cause for concern, there was an even larger gap in reading achievement between native English-speaking Hispanic students and Hispanic ELLs (Hemphill & Vanneman, 2011).

This is an especially significant concern given the numbers of ELLs nationwide, and in California in particular. According to the U.S. Census Bureau report on the 2007 American Community Survey, almost 20% of the school age population (children between the ages of 5 and 17 years old) spoke a home language other than English (Shin & Kominski, 2010), and this group has been increasing in recent years. From 2000 to

2009, the number of ELLs enrolled in preschool through twelfth grade increased by 14%, while total enrollment over that same period only increased by 5% (National Clearinghouse for English Language Acquisition; NCELA, 2010). During the 2015-2016 school year, ELLs made up about 22% of the total school-aged population in California (California Department of Education; CDE, 2016). Among those ELLs, more than 50 different primary languages were reported. However, Spanish-speaking students were by far the largest group and comprised 84% of ELLs in California (see Figure 1; CDE, 2016; NCELA, 2011). Data also indicate that, among ELLs, Hispanic students are the ethnic group most at-risk, as they have the least educated parents and are the most likely to live in poverty (Halle et al., 2012). For this reason, this study focuses on Spanish-speaking ELLs and how schools can identify and prevent problems with reading achievement.

Multi-Tiered Support Systems

Given the prevalence of reading problems among Spanish-speaking ELLs, it is critical to be able to identify struggling students and provide them with assistance. From research findings on literacy development with NESs, it is clear that the gap in reading achievement identified at the fourth-grade level (NCES, 2011) has its root in earlier difficulties. Stanovich (1986) coined the term, "Matthew Effect," to describe how initial differences in literacy acquisition led to increasingly larger achievement gaps over time, and Juel (1988) found that 88% of poor readers in first grade will continue to be poor readers in fourth grade. Instruction plays a critical role in preventing the development of serious reading problems (Fletcher, Lyon, Fuchs, & Barnes, 2007; Scanlon, Vellutino, Small, Fanuele, & Sweeney, 2005), which makes it vitally important for schools to identify struggling readers early so that targeted instruction can be provided to prevent future problems.

Multi-tiered support systems (MTSSs) are one method schools are using to meet the need both for early identification of students who are at risk for poor reading outcomes and for provision of targeted instruction to correct those problems (Fuchs & Fuchs, 2004; Gersten et al., 2008; Linan-Thompson & Vaughn, 2010). Although most research on multi-tiered support systems has been done with native English speakers (Klingner & Edwards, 2006), there is evidence that ELLs, even those with low English proficiency, can benefit from early intervention within an MTSS framework (Gersten et al., 2007; Linan-Thompson, Cirino, & Vaughn, 2007; O'Connor, Bocian, Beebe-Frankenberger, & Linklater, 2010; Vanderwood & Nam, 2007). In fact, early intervention is particularly important for ELLs, as attaining proficiency early is linked with better reading outcomes long-term (Halle et al., 2012).

Tiers within a MTSS. There are typically three tiers within a multi-tiered support system (e.g., National Association of State Directors of Special Education, 2007). The first tier, Tier 1, includes all students and the support provided is core instruction. Universal screening is conducted to identify students who are at risk for reading failure and in need of additional support (Gersten et al., 2008).

At-risk students are then given supplemental reading instruction or intervention. This is referred to as Tier 2. Tier 2 interventions are more focused and targeted than core instruction and should cover no more than one to three of the five foundational literacy skills (phonological awareness, phonics, fluency, vocabulary, and comprehension;

Gersten et al., 2008). Once students have been placed in Tier 2 intervention, their progress is then monitored on a weekly or biweekly basis to measure their rate of progress towards specific goals.

Students who do not make sufficient progress over time are then moved into Tier 3. Tier 3 interventions are even more targeted and intensive than Tier 2. Typically, if students continue to struggle at this level, they are referred for testing and possible placement in special education. Interestingly, while ELLs are overrepresented in special education, research indicates that, proportionally, the number of nonresponders among ELLs are consistent with reported findings for native English-speaking students (August & Shanahan, 2006; Linan-Thompson, Vaughn, Prater, & Cirino, 2006)

Screening assessments. One of the essential aspects of a MTSS is reliance upon data for decision-making at each tier (National Center on Response to Intervention, 2010), and one of the most important datasets is that obtained from universal screening (Clemens, Shapiro, & Thoemmes, 2011; Jenkins, Hudson, & Johnson, 2007). Universal screening allows schools to evaluate the efficacy of their core instruction. It also allows for the identification of specific students who need additional assistance. This is foundational to the entire MTSS approach as this identification process leads to the provision of intervention for the at-risk students (Gersten et al., 2008; Johnson, Jenkins, Petscher, & Catts, 2009; Kupzyk, Daly, Ihlo, & Young, 2012; NCRTI, 2010).

Literacy Development

When implementing a MTSS, it becomes critical to choose the correct skills to assess in order to identify struggling students who need intervention. Researchers have studied extensively how literacy develops and which early literacy skills can best predict later reading achievement (Mesmer & Williams, 2014; National Institute of Child Health and Human Development [NICHHD], 2000; Snow et al., 1998). Key literacy skills include phonemic awareness (PA), phonics, and fluency with reading comprehension as the end goal of literacy (NICHHD, 2000; Snow et al., 1998). First grade is a critical period for assessing literacy as students should have early literacy skills established, though most students are not yet able to fluently read connected text (Clemens et al., 2011; Mesmer & Williams, 2014; NICHHD, 2000; Snow et al., 1998).

As students develop their skills, PA is one of the first areas they must master in order to have a strong foundation for the development of phonics (NICHHD, 2000; Stanovich & Siegel, 1994; Torgesen, Wagner, Rashotte, Burgess, & Hecht, 1997; Vellutino, Tunmer, Jaccard, & Chen, 2007). PA refers to the recognition that words are made up of a series of sounds. With this knowledge in conjunction with the alphabetic principle, the awareness that sounds are represented by letters and letter combinations, a student can develop phonics, or decoding skills (Fletcher et al., 2007; Lesaux, Geva, Koda, Siegel, & Shanahan, 2008). However, since English is an orthographically irregular language, students must also learn to memorize certain “sight words” (i.e., words that do not follow the typical phonetic pattern; Gough & Walsh, 1991, as cited in Clemens, 2009). As students become more proficient with phonics and word recognition, they move from the acquisition to the fluency phase of word reading (Kame’enui & Simmons, 2001), and in early elementary, students’ oral reading fluency is strongly correlated with their reading comprehension (Kuhn & Stahl, 2003).

Literacy development for bilinguals. Researchers have found that the same early literacy skills that are important for NESs are critical for ELLs as well, and those skills develop in the same sequence for both groups (August & Shanahan, 2006; Lesaux & Siegel, 2003; Solari et al., 2014). However, there is also evidence that additional skill areas may play a role in literacy development for bilingual students. The primary areas that researchers have considered are the impact of students' oral language proficiency on their literacy development, and the role that students' early literacy skills in their first language (e.g., Spanish) play in their second language (English) literacy development.

Among NESs, almost every child possesses sufficient oral language proficiency to comprehend first-grade level texts, which have relatively low language demands (Gottardo & Mueller, 2009). For ELLs, on the other hand, oral language proficiency develops more slowly (Manis, Lindsey, & Bailey, 2004). This difference means that an ELL's level of oral language proficiency may be a significant predictor of his/her reading comprehension (August & Shanahan, 2006; Gottardo & Mueller, 2009; Kim, 2012), though some studies of Spanish-speaking ELLs have not found that connection (e.g., Mancilla-Martinez & Lesaux, 2010; Manis et al., 2004).

Similarly, there is mixed evidence of the relations between oral language proficiency and oral reading fluency (ORF). In a study of 150 Spanish-speaking ELLs in first grade, Kim (2012) found that oral language skill was not uniquely related to ORF. Solari et al. (2014) found that receptive vocabulary in kindergarten predicted ORF in first grade, but once students were in first grade, receptive vocabulary was no longer a predictor. However, as Solari and colleagues pointed out, the development of ORF in a Spanish-speaking ELL population has not been extensively studied.

In addition, researchers have considered the possibility of cross-language interactions (i.e., the impact of a student’s Spanish early literacy skills on the development of English early literacy skills and reading comprehension). While there is evidence of cross-language transfer, particularly with PA (Lindsey, Manis & Bailey, 2003; Quiroga, Lemos-Britton, Mostafapour, Abbott, & Berninger, 2002), the pattern across studies indicates that Spanish skills are not significant predictors of English literacy for students who are only receiving instruction in English (Gottardo & Mueller, 2009; Mancilla-Martinez & Lesaux, 2010). Among such students, the effect of Spanish literacy skills on reading comprehension is either mediated by the students’ skills in English (Kim, 2012; Manis et al., 2004) or becomes insignificant once the equivalent English variables are included (Manis et al., 2004; Solari et al., 2014).

Chapter 2: Review of Selected Literature

When it comes to screening instruments, curriculum-based measures (CBM) are the most common type of screening assessment. Deno proposed the use of CBM in his seminal 1985 article. CBM were initially created as short, fluency assessments that helped special education teachers to monitor students' progress in reading and to adjust instruction accordingly. While oral reading fluency is a well-established CBM for elementary students and can predict later comprehension scores, it is too difficult for most first graders at the beginning of the year, so research on other measures was needed (Fuchs, Fuchs, & Compton, 2004). Over time, additional CBM have been developed and used for screening, progress monitoring, predicting state test scores, and assessing both NES and ELLs (Deno, 2003). Prior research has found these measures to be reliable and valid for use within a MTSS (Kame'enui, 2002). As was discussed previously, these measures should be chosen to inform instruction and should be matched with the process of reading development that typically occurs at this age (Gutiérrez, 2010).

Evaluating Screening Measures

Specific factors must be considered when developing or selecting a screening tool, and given the importance of the intervention decisions involved, it is critical to use high-quality screening assessments. Screeners must be technically adequate, able to predict future performance, and able to accurately classify students as at-risk or not at-risk. They should also provide information that will guide school staff in developing appropriate interventions, and they must be simple enough to administer that a school's resources are not overburdened (Fuchs & Fuchs, 1998; Hayes, Nelson & Jarrett, 1987; Reschly & Wilson, 1995).

When examining whether a screener is technically adequate, reliability is the first consideration. Reliability refers to a measure's consistency in scores across raters, time periods, or different versions of the measure. For screening purposes, experts recommend that a test should have a reliability of at least .80 (Salvia, Ysseldyke, & Bolt, 2010).

Another critical psychometric property is an assessment's validity. In their joint publication, Standards for Educational and Psychological Testing (the Test Standards), the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) defined validity as, "the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" (AERA, APA & NCME, 1999, p. 9). As one can see in the definition, validity is specific to how a particular test is going to be used. It is a fundamental requirement that must be met before any assessment can be used in a meaningful way.

For screening assessments, certain types of validity are important, such as content validity and predictive validity. Content validity refers to whether the test appropriately samples from the knowledge domain it is designed to measure, while predictive validity refers to the relations between measures, typically between a newly designed measure and one that has already been established as a valid measure for that particular content area. An effective screener also needs to have good discriminant power (i.e., the ability to discriminate between individuals with high vs. low levels of the trait being measured).

When used with a subgroup, such as ELLs, a screener’s predictive validity needs to be examined specifically for that group (AERA et al., 1999). If a measure exhibits different predictive ability for different subgroups, for a reason that is unrelated to the construct of interest, this disparity is referred to as predictive bias (Klein & Jimerson, 2005; Salvia et al., 2010). With screening tools, one important measure of predictive validity is diagnostic accuracy.

Diagnostic accuracy refers to an assessment’s ability to predict future performance on a specified criterion measure. The two key indices of a measure’s diagnostic accuracy are sensitivity and specificity. Sensitivity is a measure of how well the predictor identifies students who fail the criterion measure, while specificity measures how well the predictor identifies students who pass the criterion measure. Both indices are based on the cut score used, that is, what score on the screener would predict success or failure on the criterion measure (Jenkins et al., 2007; Rathvon, 2004).

The lower the sensitivity level, the more truly “at-risk” students will be missed (Johnson et al., 2009). With screening tools, this translates into failure to identify students who need intervention services. On the other hand, the lower the specificity level, the more students who will be incorrectly identified as at-risk, which can overburden a school’s resources and ability to provide intervention to all at-risk students. Within a MTSS framework, the goal of screening is to identify at-risk students and provide them with intervention, so it is critical to have a high sensitivity level. Jenkins and colleagues (2007) recommended a minimum standard for sensitivity of 90 percent. However, there is no agreed upon standard for specificity, as that depends upon the school’s resources and their willingness to provide intervention even to borderline students (Clemens et al., 2011). There is a trade-off between sensitivity and specificity such that requiring high sensitivity may necessitate accepting less than ideal specificity (Glover & Albers, 2007). However, Compton et al. (2010) recommended a minimum standard for specificity of 80 percent.

Phonological Awareness

With the psychometric criteria for screening tools in mind, one can evaluate the screeners that are typically used to measure different skill areas. As mentioned previously, phonemic awareness (PA) is one of the first literacy skills a student needs to develop. Assessment of PA typically starts in kindergarten. For example, the commonly used Dynamic Indicators of Basic Early Literacy Skills (DIBELS) include Initial Sound Fluency (ISF), a measure of emerging phonological awareness, and Phoneme Segmentation Fluency (PSF), a measure that assesses the more complex skill of fluently segmenting words into their component phonemes (e.g., Burke, Hagen-Burke, Kwok, & Parker, 2009). By first grade, students are expected to move beyond segmenting the first sound of a word to fully segmenting one-syllable words, so PSF is recommended for assessing students' PA (Good & Kaminski, 2002).

Phoneme Segmentation Fluency (PSF). Although PSF is the most commonly used PA screener at the beginning of first grade, there is limited evidence of PSF's effectiveness as a screener for ELLs. The majority of PSF studies have been conducted with NESs and either did not include ELLs or did not disaggregate and report their results (e.g., Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009). This leaves a gap in the existing research.

However, when considering whether PSF should be used with ELLs, one should first note that research with NESs has found evidence that, despite its widespread use, PSF is not a good screener during first grade. The problems with PSF's ability to predict future reading performance have been attributed to issues with both reliability and validity (e.g., Clemens, Hilt-Panahon, Shapiro, & Yoon, 2012; Riedel, 2007). Researchers have repeatedly found PSF is less able to predict end-of-first-grade reading comprehension than other commonly used measures, such as Nonsense Word Fluency (NWF) and Word Identification Fluency (WIF; Clemens et al., 2011; Clemens et al., 2012; Johnson et al., 2009; Riedel, 2007). This problem is only exacerbated when researchers attempt to predict even more advanced skills, such as third grade standardized tests scores (e.g., Goffreda, DiPerna, & Pedersen, 2009; Munger & Blachman, 2013).

There are a few different factors that may contribute to this identified weakness in predictive validity. Researchers have hypothesized that the reliability of PSF is low because of the complexity of the scoring rules and because student scores are limited by the examiner's fluency with administration of the test (Clemens et al., 2012; Riedel, 2007). Any weakness in reliability would weaken the correlations between PSF and whichever criterion measure is used and result in low predictive validity.

There is also evidence that PSF may not adequately capture the variety of PA skills that students need to develop in order to read well. As the name indicates, PSF measures students' ability to segment words into individual phonemes. However, research conducted with preschoolers (Slocum, O'Connor, & Jenkins, 1993) found that training in one particular type of phonological manipulation (e.g., segmentation) did not enable students to transfer that skill to other phonological manipulation tasks. The researchers concluded that "performance on a particular phonological manipulation task should not be equated with PA" (p. 628).

This caution was borne out in research conducted with kindergartners comparing the effectiveness of three different measures of PA: initial sound fluency (ISF), PSF, and a combined phoneme segmentation task (CPST; Linklater, O'Connor, & Palardy, 2009). The researchers examined results for NESs and ELLs separately and found that PSF given in fall of kindergarten did not significantly predict later reading performance better than ISF for NESs or ELLs. Interestingly, CPST, a broader measure that incorporated elements of both ISF and PSF, was found to be the best predictor of students' end of year performance on NWF and the Woodcock Reading Mastery Test – Revised Normative Update (WRMT-R/NU) for both groups.

At the first-grade level, researchers have hypothesized that even students who struggle with reading are likely to have already mastered phonemic segmentation by the middle of the year, since it is a comparatively low-level skill (Riedel, 2007). In that case, PSF would be unable to discriminate between students with strong reading skills and those with weak ones. This inability to discriminate is one way that PSF's weak construct validity would lead to low predictive validity.

Despite these weaknesses, PSF continues to be widely used. Given that widespread use and the need to support ELLs in their literacy development, it is still important to research PSF and its effectiveness as a predictor for the ELL population specifically. There are a few researchers who have done so.

In a 2007 study, Riedel examined the relations between four DIBELS measures (Letter Naming Fluency, LNF; PSF; NWF; and ORF) administered in fall, winter, and spring, and end-of-year scores on the Group Reading Assessment and Diagnostic Evaluation (GRADE) in first grade and the TerraNova Reading subtest in second grade. Within his larger sample of 1,518 first grade students, a small number of students (*n* = 59) were ELLs. Riedel noted that "English-reading ability varied substantially across these students" (p. 551), so he chose to analyze their results separately. However, due to the small size of the sample, ELLs were excluded from the receiver operating characteristic (ROC) and logistic regression analyses in the study, and only Pearson correlations were calculated for them.

Among NESs, Riedel found that PSF was the weakest of the four predictors. In fact, he noted that winter and spring PSF scores predicted first- and second-grade reading comprehension at a rate that was only slightly better than chance (53-58% classified correctly). However, as mentioned previously, ELLs were not included in this analysis.

For ELLs, fall PSF was more strongly correlated with the GRADE than winter and spring PSF were (.31, .19, and .21, respectively), and at each period, NWF's correlations with the GRADE were higher (.41, .47, and .41). Analysis of the NES sample found similar correlations for PSF (.26, .16, and .15) although the author made no mention of whether these results were significantly different.

Unfortunately, given the small sample size and the limited analyses, not much can be concluded for ELLs from this study. The correlations found were weak which raises a red flag, but there were no data on what cut scores should be used or how accurately PSF could predict for ELLs, the primary concern when using screening tools. Fortunately, other researchers have investigated this in more detail.

Johnson, Jenkins, Petscher, and Catts (2009) examined end of kindergarten and beginning of first grade screening scores for approximately 12,000 students to see which best predicted end-of-first-grade Stanford Achievement Test (SAT) scores. They were specifically interested in validating cut scores and making sure that they held true for the ELL population. A subsample (*n* = 1,217) of ELLs was created and used to identify the strongest predictors and the specific cut scores needed to reach 90 percent sensitivity.

When the researchers analyzed students' scores from the beginning of first grade, they found that ORF scores predicted end-of-year performance better than either PSF or NWF, and this held true for both ELLs and the larger sample. However, the ELL group required lower cut scores (-5 points) to achieve 90 percent sensitivity. While this may only be a small difference practically speaking, the authors concluded that schools should disaggregate their screening results, especially for ELLs, to make sure that they have equal access to Tier 2 interventions and are not excluded inappropriately. Nonetheless, they did not report findings for ELL students on PSF specifically, leaving one to wonder whether this measure should be handled similarly.

Vanderwood, Nam, and Sun (2014) examined first graders' fall scores on LNF and fall, winter, and spring scores on PSF and NWF to determine how strongly they correlated with winter and spring ORF and Maze (a group-administered screener that uses a multiple-choice, cloze task to measure reading comprehension). They intentionally selected a sample of only Korean-speaking ELLs (*N* = 30). Unfortunately, like Riedel (2007), Vanderwood and colleagues only ran correlational analyses, most likely due to the small sample size. The size of the sample also prevented them from differentiating the students by English proficiency.

They found that PSF was not significantly correlated with any of the other measures given (correlations ranged from -.22 to .29). By the end of the study, all students in the sample scored in the “low risk” range on ORF (based on DIBELS benchmarks). However, several students (7 out of 30) were still scoring in the “emerging” range on PSF as categorized by DIBELS benchmarks.

Unfortunately, one can only draw limited conclusions from this study. The sample was confined to Korean-speaking ELLs and none of the students in the sample scored in the “at-risk” range on end-of-year ORF. Obviously, there is no way to know whether findings would be similar for Spanish-speaking ELLs (or other language groups), particularly if those students were struggling readers. In addition, neither ROC or regression analyses were run, so there is no data on how different cut scores might influence the effectiveness of PSF as a screening tool.

Fortunately, Vanderwood’s research team conducted an additional study. Han, Vanderwood, and Lee (2015) analyzed winter of first grade PSF, NWF, and WIF scores for 102 Korean-speaking ELLs. They examined how well those scores could predict spring of first grade scores on ORF and the Woodcock Reading Mastery Tests – Revised

(WRMT-R). With this sample, they ran regression and ROC analyses in addition to correlations, and they also grouped students into three different categories based on their English language proficiency as measured by California English Language Development Test (CELDT) scores: Beginning/Early Intermediate (*n* = 27), Intermediate (*n* = 42), and Early Advanced/Advanced (*n* = 33).

Unlike the prior study, they found that PSF was significantly correlated with ORF and the WRMT (.35 and .34, respectively). However, the strength of the correlation varied by level of English proficiency. Correlations were stronger for the Beginning/Early Intermediate (.51 and .40) and Intermediate (.44 and .37) groups and were weakest for the highest-level group (-.20 and -.23). In fact, for the highest group, correlations were no longer significant, and the researchers noted that many students who scored below the PSF benchmark were considered fluent readers based on ORF scores. Given this weakness in the relations between PSF and ORF for the Early Advanced/Advanced group, it is unsurprising that ROC analyses of the entire sample of Korean-speaking ELLs found PSF was unable to accurately predict ORF outcomes based on either the DIBELS recommended cut score (35, sensitivity = .73, specificity = .71) or the optimal cut score calculated by the researchers (28, sensitivity = .73, specificity = .85). The hierarchical regression analyses found that PSF explained less variance than either of the other screeners (.11 on both ORF and WRMT, compared with .22 and .19 for NWF and .42 and .30 for WIF).

Of the studies conducted with ELLs, this is the only one that considered level of English proficiency, and they found that it impacted PSF's predictive power for Korean-speaking ELLs. Since Korean-speaking ELLs are not a large portion of the total ELL population, there is still a need to conduct similar studies with Spanish-speaking ELLs in particular, as well as with other language groups. In conducting such studies, researchers should be mindful of this finding and examine whether ELLs' level of English proficiency impacts the relations between PSF and the outcome variable used.

Summary. Studies with NESs, as well as the limited number of studies that have examined ELLs' scores on PSF, agree that there are potential problems with this measure. Researchers have raised concerns about reliability, content validity, and predictive validity, though many practitioners continue to rely on this assessment. The proposed study would include PSF as a screening measure for ELLs in first grade and would address some of the concerns other researchers have raised.

Predictive validity is probably the most important issue for a screener, so weakness there is a noteworthy concern. However, this concern can be addressed, in part, by considering the literacy development process and the intended use of early literacy screeners. As O'Connor and Jenkins (1999) pointed out, "The relative accuracy of prediction varies with the specific measures used as predictors and as outcomes, the timing of their administration, and the degree and direction of classification error the researchers consider acceptable…" (p. 160). This is particularly important for a screening tool like PSF which was designed to measure a skill that is only one step within the literacy development process. PA is a necessary skill for developing reading, but it is not sufficient for reading comprehension by itself. Rather, PA "provides the foundational skills that will make later acquisition of decoding skills and subsequent reading fluency and comprehension more likely” (Burke & Hagen-Burke, 2007, p. 75).

With that framework in mind, it is unsurprising that researchers have found PSF is less able to predict end-of-first-grade reading comprehension than other measures, such as NWF and WIF, which measure different skills that are farther along in the literacy development process. This problem is only exacerbated when researchers attempt to predict even more advanced skills, such as standardized test scores in later grades. In this case, PSF’s weak predictive validity may be an issue with the measure itself or may simply be an artifact of attempts to extend PSF’s predictive power beyond what is possible given the construct it measures and PA’s place within the literacy development process.

For this reason, the proposed study would consider PSF’s predictive power within a developmental framework, specifically, how well can PSF given in fall of first grade predict students’ phonics skills in winter of first grade? Such a question fits with the intended use of early literacy screeners and matches how screening tests are used in school settings within an RTI framework. In addition, to address reliability, the proposed study would use examiners who are well-trained to administer early literacy screeners within an RTI setting.

Construct validity is an issue that is more difficult to address. PSF is a narrow measure of one particular skill, phonemic segmentation, and the examiner presents all information orally. Segmentation may not be a broad enough measure of PA skills (Slocum et al., 1993), and as others have noted, PSF may be a weaker predictor than measures such as NWF and WIF because it is less like actual reading (Clemens et al., 2012; Riedel, 2007). For this reason, the proposed study would compare PSF with the Developmental Spelling Test (DST; Tangel & Blachman, 1992), a measure that evaluates students' developing ability to write and spell the phonemes in words.

Spelling. The development of spelling skills overlaps with reading development (Chua, Liow, & Yeong, 2016; Clemens, Oslund, Simmons, & Simmons, 2014; Harrison et al., 2016), and spelling can be thought of as "encoding", the opposite of decoding. It has also been described as "applied phonological skill" (Lesaux & Geva, 2006) since spelling requires PA skills and integration with orthographic knowledge (Chua et al., 2016). Researchers have found relations between PA and spelling for Spanish-speaking ELLs, specifically (Ford, Cabell, Konold, Invernizzi, & Gartland, 2013; Goodrich, Farrington, & Lonigan, 2016; Harrison et al., 2016). However, it is only recently that researchers have started examining whether spelling could be used to predict future literacy skills (Chua et al., 2016).

Ford, Cabell, Konold, Invernizzi, and Gartland (2013) used cluster analysis to examine fall of kindergarten scores on the Phonological Awareness Literacy screening for kindergarten (PALS-K) which included measures of PA, alphabet knowledge, and phonetic spelling. Their sample consisted of 2,351 Spanish-speaking ELLs receiving ESL services. They found four distinct profiles that were associated with PALS literacy outcomes in spring of K and fall of 1st grade. Interestingly, they observed that the two profiles that were associated with the strongest performance on the outcome measures were those that did better on orthographic skills (i.e., alphabet knowledge and phonetic spelling) in kindergarten. They also noted that two of the cluster profiles included students with adequate PA skills at the beginning of kindergarten, but only the group that also had strong orthographic skills ended up scoring well on the outcome measures. This led the researchers to conclude that, “While PA may be a necessary precursor to reading, PA in the absence of orthographic skills may not be sufficient.” (p. 889)

Morris and colleagues (2017) explored a similar issue using a sample of NESs and a more typical analysis of predictive validity. They compared the DIBELS early literacy screeners (PSF, NWF and ORF) with alternative tasks designed to measure the same skill area. They compared: PSF with a measure of phonemic spelling (phSPEL); NWF with a word reading task, word recognition-timed (WR-t); and ORF with an oral reading accuracy and comprehension task, graded passage reading (grPASS). For their study, they assessed two different cohorts of students (*N* = 319) at multiples times from kindergarten through third grade (though PSF and phSPEL were only assessed in winter and spring of kindergarten and fall of first grade.)

They found that, in both cohorts, phSPEL was moderately correlated (.48-.63) with word reading scores at the next screening period (NWF and WR-t, respectively) and outperformed PSF which was weakly to moderately correlated (.24-.48) with those same measures. In first grade specifically, the fall-to-winter PSF to NWF prediction coefficient was .34, while the phSPEL to NWF prediction coefficient was .48. From these results, they concluded that phonetic spelling was a better measure than PSF and that, “the added letter/sound component of phSPEL gave it an advantage over PSF, a pure oral segmentation task” (p. 309). They also highlighted the way that spelling slows down the process of segmentation for students and gives examiners time to score student responses. While the researchers acknowledged that phSPEL was somewhat difficult to score given the necessary rules for determining which letter responses could be considered correct for each phoneme, they also pointed out that examiners have time to do this scoring afterward, whereas PSF requires examiners to score students' responses as quickly as possible during administration.

Clemens, Oslund, Simmons, and Simmons (2014) looked specifically at how different methods of scoring spelling impacted the test's concurrent (end of kindergarten) and predictive (end of first grade) validity. They compared five different scoring methods: total correct words, total correct letter sounds, total correct letter sequences, a rubric for invented spellings (DST), and calculation of a Spelling Sensitivity Score (SSS). Their sample consisted of 287 NESs who were identified as at-risk at the beginning of kindergarten as part of a larger longitudinal study. They used a latent variable model to compare spelling scores with measures of PA, pseudoword decoding, word reading, and first-grade reading skills. Three measures were used to assess PA: PSF and two subtests (Blending Words and Sound Matching) from the Comprehensive Test of Phonological Processing (CTOPP). Pseudoword decoding was also assessed with three different measures: NWF, the Word Attack subtest from the Woodcock Reading Mastery Test (WRMT-R/NU), and the Phonemic Decoding Efficiency subtest from the Test of Word Reading Efficiency (TOWRE). Word reading was assessed with the Word Identification subtest from the WRMT-R/NU, the Sight Word Efficiency subtest from the TOWRE, and ORF. Lastly, first-grade reading skills included ORF, the Word Identification and Word

Attack subtests mentioned previously, and the Passage Comprehension subtest from the WRMT-R/NU.

They found that end of kindergarten spelling explained a considerable proportion of the variance in first grade reading skills (from 39% to 46%) and explained unique variance over and above kindergarten word reading skills. They also noted that the four measures that allowed for partial or invented spellings (correct letter sounds, correct letter sequences, rubric, and SSS) explained more variance than conventional scoring (correct words). The rubric scoring method (DST) which was designed to measure phonemic spelling and incorporated a scaled scoring system (Tangel & Blachman, 1992) accounted for the most variance in PA skills compared to the other spelling measures, though all the spelling strategies showed similar results overall. Like Morris and colleagues, the researchers noted that the scoring rules for DST were rather complex, but they also noted that the developers (Tangel & Blachman, 1992) reported good inter-rater reliability regardless. Though their results may only generalize to similar NESs identified as at-risk when starting school, they concluded that spelling assessment is underutilized in evaluating early reading skills.

Chua and colleagues (2016) examined spelling as a predictor for 127 bilingual (Mandarin and English) kindergartners. They used tests of five different early literacy skills (PA, vocab, spelling, letter identification, and rapid automatized naming) in winter of kindergarten to predict spring word reading below the 25th percentile. They analyzed the data using logistic regression. While the researchers did not differentiate among ELLs with different levels of English proficiency, they did include two types of bilingual students – those for whom English was their first language (NESs), and those for whom Mandarin was their first language (ELLs).

They found that spelling scores were the single best predictor (sensitivity of .75 and specificity of .73 with cut score selected to maximize these values). While spelling was a better predictor than the PA test, the full battery still predicted better than any of the individual measures with a sensitivity of .81 and specificity of .87. They concluded that spelling was a stronger measure because it required a broader range of skills. Like early reading, strong spelling skills require the integration of PA and phoneme segmentation with letter-sound knowledge.

Summary. Overall, while the research is limited, there is evidence that spelling holds promise as a strong predictor of later reading skills since it could potentially capture variance in both PA and orthographic skills. Spelling tests are also relatively easy to administer and score, an important criterion for screeners. However, at this point, it has not been researched as a predictor for Spanish-speaking ELLs. Though there are different ways of scoring spelling tests, there is some evidence that DST can capture more variance than other scoring strategies (Clemens et al., 2014) due to the creators' intentional focus on the development of spelling skills (Tangel & Blachman, 1992). For this reason, the proposed study would include DST as a first-grade screener and compare its predictive power to that of PSF and other early literacy screeners.

Word Reading

Researchers have examined two measures of phonics skills for use with first graders. Nonsense Word Fluency (NWF) is a screener used to assess a student's phonics ability by asking him/her to decode nonsense words (pseudowords). The choice to use nonsense words removes the possibility that students may have already seen and memorized some of the words (Nam, 2011). The words created for this measure follow a vowel-consonant (VC) or consonant-vowel-consonant (CVC) pattern, and students must use the most common pronunciation for each letter (as specified in the manual; Good & Kaminski, 2002) in order to receive credit. Word Identification Fluency (WIF), on the other hand, is a broader measure. To do well on this screener, students need to demonstrate both phonics and orthographic skills since WIF includes real words with a larger variety of possible letter pronunciations as well as sight words, which are not decodable (Fuchs et al., 2004). Like NWF, it has been used to assess first graders' word reading skills and predict future reading performance. WIF specifically measures a student's ability to read grade-appropriate, high-frequency words quickly and accurately (Deno, Mirkin, & Chiang, 1982; Fuchs et al., 2004).

Researchers have examined the technical aspects of both NWF and WIF and evaluated them for use as screening instruments. While research has been done on both measures, there is still some debate as to when the related skills develop and which one is a better screener (e.g., Aaron et al., 1999; Compton, 2000). In the following sections, I will describe the current state of the research on NWF and WIF and evaluate how well they meet the criteria for good screening instruments.

Nonsense Word Fluency (NWF). Catts, Petscher, Schatschneider, Bridges, and Mendoza (2009) reviewed several DIBELS measures and examined the impact of floor effects on classification accuracy. Their results using a sample of approximately 17,000 first graders indicated that NWF is an acceptable screener with limited floor effects by the beginning of first grade. With end of third grade ORF as the outcome measure and sensitivity held at .90, specificity ranged from .44 to .54 across four administrations of NWF throughout the first-grade year. They also calculated the AUC for each administration of NWF, and values ranged from .84 to .88, which would be considered useful discrimination under the guidelines offered by Swets (1988).

Vanderwood, Linklater, and Healy (2008) conducted a study with 134 ELLs examining the relations between NWF in first grade and reading outcomes at the end of third grade. They found that NWF explained a significant portion of the variance in reading outcomes above and beyond that explained by language proficiency (ΔR^2 = .32, .09, and .08 on oral reading fluency, reading maze, and California Achievement Test, Sixth Edition, respectively). However, they also found that the sensitivity of the NWF cut scores previously established with an NES population was low (.43-.55 depending on the outcome measure) for this sample. Their analysis revealed that over 80 percent of the false negatives identified were students at the lowest level of English proficiency. These results suggest that, for students who are still acquiring English proficiency, a different screening measure or different cut scores may be needed to identify those truly at risk and in need of early intervention.

Fien and colleagues (2008) conducted a study examining the use of NWF as a universal screener with kindergarten through second grade students in Reading First schools. They included a large sample of both ELLs and NESs. They found that NWF explained a large amount of the variance on the criterion measures (ORF and the Stanford

Achievement Test Series, 10th edition; SAT-10) for all students, and that it predicted performance equally well for ELLs as for NESs. However, they did not examine ELLs' results at different levels of English proficiency, so differences in correlations by proficiency level may have existed but gone unidentified. The researchers did not conduct an analysis of the measure's classification accuracy either, despite its importance in a screening tool.

Jenkins, Hudson, and Johnson (2007) conducted a review of screening studies with elementary students. Although they noted that differences across studies made it difficult to identify patterns at the first-grade level, they did find that NWF showed poor sensitivity (at best, .67) when predicting the bottom quartile on the SAT-10. Although they did not specifically review studies with ELLs, this finding seems to be in line with the concerns raised by Vanderwood et al.'s 2008 study.

Johnson, Jenkins, Petscher, and Catts (2009) used a large sample (N = 12,055) of first grade students to examine the classification accuracy of several different CBM measures, including NWF. When comparing beginning of first grade PSF, NWF, and ORF scores, they found that ORF best predicted end of year SAT scores. They also found that, with a sensitivity of .90 (as recommended by Jenkins et al., 2007), specificity was poor for all measures, and for NWF was only .42 when unsatisfactory reading was defined as a SAT score below the 40th percentile.

The researchers examined results for ELLs as well. They found that a lower cut score was needed to reach 90% sensitivity for this subgroup. From this finding, they concluded that schools need to disaggregate results and set appropriate cut scores for

ELLs in order for them to have equal access to Tier 2 interventions. However, they did not go further and analyze ELLs' results by English language proficiency level.

Word Identification Fluency (WIF). Although WIF is not as widely used as the popular AIMSweb or DIBELS early literacy assessment batteries, there is evidence that suggests it is a more accurate screening measure at the first-grade level than NWF (Clemens et al., 2011; Fuchs et al., 2004; Healy, 2007). While NWF measures students' ability to quickly decode "words" they have never seen before, WIF measures a broader range of phonics skills and is influenced by other underlying literacy skills, such as vocabulary size and knowledge of orthographic patterns (Clemens et al., 2011). Though this makes WIF less useful as a tool to diagnose weaknesses in specific phonics skills, it has been found to be an advantage for screening purposes as WIF is better able to identify at-risk students and predict later reading performance (Clemens et al., 2011; Healy, 2007).

Fuchs, Fuchs, and Compton (2004) conducted a study with first-grade students specifically examining the question of whether NWF or WIF better predicted performance on measures of word identification, phonics, fluency, and comprehension at the end of first grade. They assessed a sample of 151 at-risk students identified from the participants within a larger intervention study. Assessments were administered in fall and spring of first grade, and progress monitoring assessments were conducted at least once a week for 20 weeks. They used assessment data to measure students' initial level as well as to calculate their rate of improvement.

The researchers found that WIF had significantly higher concurrent validity than NWF with almost every outcome measure. In fact, even with the phonics measure which was itself very similar to NWF (WRMT – Word Attack subtest), NWF did not outperform WIF. A similar pattern of results was found when calculating predictive validity. Overall, WIF was found to be more highly correlated with later reading outcomes. However, while it is important to know how well a screening tool predicts for at-risk students, screeners are used with the entire school population. In this study, only a restricted sample was included. Also, although some ELLs (17%) were included in the sample, their results were not analyzed separately from the NESs, so no conclusions can be drawn from this study as to whether this measure is appropriate for the ELL subgroup.

Clemens, Shapiro, and Thoemmes (2011) conducted a similar study with a sample of 138 first grade students. They administered multiple screeners in fall of first grade: Letter Naming Fluency (LNF), PSF, NWF, and WIF. They also used multiple outcome measures as the criterion – two CBM screeners (ORF and Maze) and the Test of Word Reading Efficiency (TOWRE). The criterion measures were all administered in spring of first grade. In addition, they combined scores on the outcome measure into a latent variable composite for use in analyses. The researchers compared the classification accuracy of each screening measure individually as well as in combination with the other measures. Based on ROC curve analyses, they found that WIF was the best, single-measure predictor, although specificity was improved (i.e., fewer false positives were identified) when PSF was added as an additional predictor of ORF scores. As with Fuchs, Fuchs, and Compton (2004), a small number of ELLs were included in the sample (5%), but their results were not analyzed separately from those of NESs. Unlike Fuchs, Fuchs, and Compton (2004), Clemens and colleagues did not use an at-risk sample, nor did they include any students who were in special education.

In a dissertation study, Healy (2007) examined WIF and NWF to determine whether these measures predicted differently for ELLs versus NESs. She conducted her study with a sample of 435 first grade students, of whom 34% were ELLs. A small percentage (6%) were also receiving a literacy intervention during the screening study. WIF and NWF were used as the predictors, and students were assessed in fall, winter, and spring of first grade. ORF was used as the primary outcome measure and was administered in spring. In addition, a smaller subsample (100 students) was assessed with the WRMT-R in spring.

Healy found that WIF was a better predictor of end-of-the-year reading ability than NWF for both ELLs and NESs. She found that incorporating rate of growth into the model along with initial level explained additional variance. In this study, ELL status was not found to account for significant variance in student reading achievement. However, ELLs were treated as a homogeneous group, without being differentiated according to the extent of their English language proficiency. Overall, this study was an important addition to the research literature on WIF as the Clemens et al. study (2011) did not include any ELLs in the participants, and Fuchs et al. study (2004) only included a small percentage (17%) which were not analyzed separately.

Compton, Fuchs, Fuchs, and Bryant (2006) examined WIF's utility from a slightly different perspective. Specifically, they researched whether including WIF (initial level and/or growth over time) improved predictive power when added to a screening battery that already included phonemic awareness, rapid naming, and oral vocabulary measures. They found that adding initial WIF scores using a direct route approach to screening did not improve classification accuracy. However, when they added 5 weeks of progress monitoring with the WIF measure for those students identified as at-risk, the additional WIF-Level and WIF-Slope data significantly improved the accuracy of their prediction model.

Compton and colleagues extended their team's prior research in a 2010 study of a first-grade screening battery and the application of a multiple-gating procedure to decrease the number of false positives identified. First, they compared the utility of several different, standardized, word-level reading measures (word identification, word attack, sight word efficiency, and phonemic decoding efficiency) as the initial step in the gated screening procedure. Then, they compared short-term WIF progress monitoring (intercept and slope) data, the results of a dynamic assessment teaching three different patterns of pseudoword decoding (CVC, CVC*e*, and CVC(C)*ing*), as well as running records and oral reading fluency scores as additional screening measures in contrasting models. The researchers found that, for the initial step in the gating process, the measure of phonemic decoding efficiency significantly reduced the number of students who required further assessment. At the second step, both WIF progress monitoring and the dynamic assessment of decoding skills significantly decreased the number of false positives identified.

Zumeta, Compton, and Fuchs (2012) researched how variations in the words chosen for the WIF measure impacted its usefulness as a screener. From a pool of 704 first grade students, they selected a representative sample of 204 students as well as a low achieving subgroup (*n* = 202) and an average/high achieving subgroup (*n* = 213). They compared a “broad” WIF measure where included words were sampled from a pool of 500 high-frequency words with the “narrow” WIF measure used in other studies. (This standard WIF measure includes words sampled from a pool of only 133 words, those found on the Dolch word lists for preprimer, primer, and first grade.) The two WIF measures were administered in fall of first grade, and 17 weeks of progress monitoring data were collected as well, which allowed the researchers to examine both level and rate of growth. They used multiple outcome/criterion measures administered in spring of first grade – the Word Identification and Word Attack subtests from the WRMT-R, reading fluency on a passage from the Comprehensive Reading Assessment Battery (CRAB), and the Decoding and Sight Word sections of the TOWRE.

Interestingly, they found an interaction effect between measure breadth and achievement level. The narrow WIF measure was better for screening the representative group and the average/high achieving subgroup, while the broad measure was more accurate for the low-achieving subgroup. In addition, they found that the broad WIF screener was also a better progress monitoring tool for all groups.

Summary. Overall, the research on NWF and WIF indicates some weaknesses in the measures. NWF has been found to have poor sensitivity, well below the 80% standard recommended by Compton and colleagues (2010). This means that the measure is over-identifying students as at-risk, and in schools with large numbers of truly at-risk students, this over-identification places an additional burden on a school which may already be struggling with resource issues.

Research with ELLs indicates that, while NWF predicts performance for ELLs in addition to NESs (Fien et al., 2008), there are differences. Researchers who examined classification accuracy found that previously established cut scores were not as sensitive with the ELL population as with NESs (Johnson et al., 2009; Vanderwood et al., 2008). For this reason, different cut scores may be needed to ensure equal access to resources.

Research with WIF indicates that it is a stronger measure than NWF overall. However, very little research has been done with ELLs (Fuchs et al., 2004; Healy, 2007) and none has considered the potential impact of differences in ELLs' English language proficiency. It is possible that the high frequency words used in the WIF measure are less familiar to ELLs than NESs. This difference could change the skill being measured and impact the screener's predictive validity (Healy, 2007).

Purpose of Study

The purpose of this study was to examine the predictive power and diagnostic accuracy of four early literacy screening measures with Spanish-speaking ELLs: PSF, DST, NWF, and WIF. The classification accuracy of each measure was examined and compared to see how these screeners differ in their predictive power for students with different levels of English proficiency. Specifically, the following research questions guided the investigation:

1. What are the relations between the early literacy screening measures (PSF, DST, NWF, and WIF) collected in fall, winter, and spring of first grade and the outcome variable, spring ORF?
    a. To what extent do the relations between variables differ for students with different levels of English language proficiency, as measured by CELDT scores?
2. How much variance in reading outcomes (spring ORF) is explained by the early literacy predictor variables (PSF, DST, NWF, and WIF) administered in fall and winter?
    a. To what extent does English language proficiency, as measured by CELDT scores, explain additional variance in outcomes after controlling for student performance on early literacy measures?
    b. To what extent does one aspect of English language proficiency, receptive vocabulary as measured by PPVT scores, explain additional variance in outcomes after controlling for student performance on early literacy measures?
3. Considering screeners from within a literacy development framework, which winter early literacy measure explains the most variance in spring ORF?
    a. Which of the fall early literacy measures accounts for the most variance in that winter predictor?

4. What is the accuracy (area under the curve, sensitivity, and specificity) of PSF, DST, NWF, and WIF in predicting scores below the 25th percentile on spring ORF?
    a. Which cut scores provide the best combination of sensitivity and specificity for screening purposes?
    b. To what extent is there a difference in predictive accuracy for students with differing English proficiency levels as measured by the CELDT?
    c. Is there a practical difference in cut scores obtained for students with differing English proficiency levels as measured by the CELDT?

Chapter 3: Methods

Participants and Setting

The data used in this study were originally collected as part of a larger, longitudinal study of the effectiveness of a response-to-intervention system (Beach & O'Connor, 2015; O'Connor, Bocian, Sanchez, & Beach, 2014). Two school districts in southern California participated in the study. The participating districts nominated school sites, and the researchers, school principals, and teachers worked together to make the final decision as to which schools would participate in the study. Five schools were selected for participation (three from one district and two from the other).

Three of the five schools were similar in size, serving approximately 450 students, while the other two schools were larger, one serving approximately 600 students and the other approximately 900 (CDE, 2017). Across schools, the percentage of students who were ELLS ranged from 33% to 60% of the total student population, and of those ELLs, approximately 95% spoke Spanish as their first language. At each school, most students (approximately 85%) identified as ethnic minorities. In District A, the largest ethnic subgroup was Hispanic students (68%) followed by African American (16%) and White (11%). Students from other ethnicities (American Indian or Alaska Native, Asian, Filipino, or Pacific Islander) each made up less than 2% of the total student population, with students of multiple ethnicities or those who chose not to respond made up the remaining 3%. In District B, the largest ethnic subgroup was Hispanic students (72%) followed by White (15%), African American (4%), and Asian (3%). Students from other ethnicities (American Indian or Alaska Native, Filipino, or Pacific Islander) each made up less than 2% of the total student population, with students of multiple ethnicities or those who chose not to respond made up the remaining 3%. Children who received free or reduced-price meals made up 52% to 95% of the population in participating schools. Table 1 lists demographic information on ELL status, ethnicity, and free and reduced lunch status for each district and its participating schools.

Core instruction and intervention. At each of the five schools, all the first-grade teachers participated in the original study. The teachers used the same reading curriculum (Houghton Mifflin Reading California), and they all participated in 120 hours of language arts professional development from the California Reading Development Center on implementing this curriculum (40 hours of training followed by 80 hours of follow-up). All students were assessed three times per year with multiple early literacy screeners. When students were identified as at-risk based on their screening scores, they were placed in Tier 2 interventions and continued to receive intervention until their assessment scores met prespecified exit criteria.

Students from the larger dataset were included in the sample for this study if they were in first grade at the beginning of the study, and they had complete data on all relevant predictors and outcomes. Applying these criteria yielded a sample of 209 first graders. English language learners (ELLs) made up 51% of the sample (*n* = 106), and based on their CELDT scores, they were at four different levels of English proficiency: Beginning (*n* = 10), Early Intermediate (*n* = 45), Intermediate (*n* = 40), and Early Advanced (*n* = 11). Fourteen percent of the students included in this study (*n* = 30) participated in intervention.

Measures

Students' skills were assessed with four early literacy screeners in fall, winter, and spring of first grade: PSF, DST, NWF, and WIF. Students' ability to read connected text was measured with DIBELS ORF in winter and spring of first grade. English proficiency was measured at the beginning of the year with the CELDT, a required test for ELLs in California, and the PPVT, a measure of receptive vocabulary.

Phoneme Segmentation Fluency (PSF). PSF is a standardized test of PA that is individually administered. It measures students' ability to segment three- or four-phoneme words into individual phonemes (e.g., *mop* = /m/ /o/ /p/). The examiner presents each word orally, and the student verbally segments the word in response. As the student responds, the examiner scores the response and then presents the next word. A student's total score is the number of correct segments produced within 1 minute. In research conducted by the test developers, the alternate-form reliability for first grade students was .60 over a period of two weeks (Kaminski & Good, 1996), and in another study, it was .67 over a period of one month (Good et al., 2004). The concurrent, criterion-related validity of PSF compared with the Woodcock-Johnson Psycho-Educational Battery readiness cluster standard score was .51 at the beginning of first grade (Good et al., 2004). The predictive validity of PSF in the fall of first grade was .54 for winter of first grade NWF and .54 for spring of first grade ORF (Good et al., 2004).

Developmental Spelling Test (DST). The DST measures students' spelling ability. An examiner can administer DST to an entire group at once, and the test is untimed. The examiner dictates each word and students write their response. With traditional spelling tests, words are simply scored as correctly or incorrectly spelled. The DST, on the other hand, gives points for incorrect spellings that are phonologically similar to the correct spelling. Specifically, each word is scored on a 7-point scale (0-6). A written response is given a score of 0 for a random or alphabetical string of letters, 1 point for a phonetically related letter, 2 points for a correct first letter, 3 points for more than one correct phoneme but not all phonemes represented, 4 points for representing all phonemes with correct or phonetically related letters, 5 points for correctly spelling all consonants and using phonetically related letters for the vowel sounds, and 6 points for a correct spelling. The students were tested with 8 words in the fall, 10 in the winter, and 12 in the spring. The test developers reported interrater reliability of .98 (Tangel & Blachman, 1992), and Clemens, Oslund, Simmons, and Simmons (2014) reported that Cronbach's alpha was .93.

Nonsense Word Fluency (NWF). As mentioned previously, the NWF screener is used to assess a student's phonics ability by asking him/her to sound out nonsense words. The test is administered one-on-one. The examiner presents the student with a page of words which follow a vowel-consonant (VC) or consonant-vowel-consonant (CVC) pattern, and informs the student that he/she may sound out the words one letter at a time (e.g., /s/ /i/ /m/) or read each one as a complete word (e.g., /sim/). The student's score is the number of correct letter sounds he/she produces in one minute (Good & Kaminski, 2002). In research conducted by the test developers, the one-month, alternate-form reliability for first grade students was .83 (Good et al., 2004). The concurrent, criterion-related validity of NWF compared with the Woodcock-Johnson Psycho-Educational

Battery readiness cluster standard score was .51 at the beginning of first grade (Good et al., 2004). The predictive validity of NWF in the winter of first grade was .82 for spring of first grade ORF (Good & Kaminski, 2002).

Word Identification Fluency (WIF). WIF is a measure of phonics that uses real words rather than nonsense words (i.e., NWF). It can also be thought of as a measure of reading fluency that uses individual words rather than connected text (i.e., ORF). WIF probes consist of 5 columns of 20 words on a single page. The words included were randomly sampled with replacement from a pool of 133 high-frequency words from the Dolch pre-primer, primer, and first-grade level lists (Fuchs et al., 2004). The examiner administers the test to the student one-on-one. As with other CBM measures, the student has one minute to read, and his/her score is the number of words read correctly (Fuchs & Fuchs, 2007). Using alternate forms, test-retest reliability after two weeks was .97, and .91 after two months (Fuchs et al., 2004), and Compton, Fuchs, Fuchs, and Bryant (2006) reported that split-half reliability exceeded .95 with their sample of low-performing first graders. The concurrent validity with the WRMT-R Word Identification subtest was .77 in the fall of first grade (Compton et al., 2006).

Oral Reading Fluency (ORF). DIBELS ORF passages measure reading rate and accuracy. The examiner tests each student one-on-one using three different passages. For each passage, the student reads aloud for one minute, and his/her score is the number of words read correctly. The median score is used for analyses and determining risk level (Good & Kaminski, 2002). According to the test developers, test-retest reliabilities ranged from .92 to .97 for elementary students, and alternate-form reliability for passages drawn from the same grade level ranged from .89 to .94 (Good & Kaminski, 2002).

English language proficiency. In this study, students' English language proficiency was assessed with two different measures.

California English Language Development Test (CELDT). At the first-grade level, the CELDT measures a student's English language listening and speaking skills. CELDT test scores are used to rate students' level of English proficiency on a 5-point scale: Beginning, Early Intermediate, Intermediate, Early Advanced, and Advanced. Students who score in the first four levels are considered ELLs, while those who score in the fifth (i.e., Advanced) category are classified as Fluent English Proficient (FEP). The CELDT is also used on an annual basis to measure students' progress in increasing their English language skills and determining whether a student should be ranked as Reclassified English Proficient (RFEP; CDE, 2003). According to the 2008 technical report from the test developers, the internal consistency as measured by Cronbach's alpha was .79 for the Listening domain and .89 for the Speaking domain at the first-grade level (CTB/McGraw-Hill LLC, 2008).

Peabody Picture Vocabulary Test-Third Edition (PPVT). The PPVT (Dunn & Dunn, 1997) measures students' receptive vocabulary in English. It is a norm-referenced assessment that is designed for use with individuals from 2.5 years old through adulthood. The examiner administers the test one-on-one. For each item, the student selects from among four pictures the one which best represents the word read by the examiner. Students' raw scores are standardized using their age in years and months at the time of testing and reported as standard quotient scores with a mean of 100 and standard deviation of 15. The test publisher reports median alternate-form reliability of .94, median test-retest reliability of .92, and median split-half reliability of .94 (Pearson Clinical, 2017). Concurrent validity between the PPVT-III and a measure of verbal ability, the Wechsler Intelligence Scale for Children-Third Edition (WISC-III), was .91 (Pearson Clinical, 2017).

Procedures

Data for this study were obtained from the whole class datasets collected during the RTI study described previously. Each screening measure (PSF, DST, NWF, and WIF) was administered three times during the academic year: fall, winter, and spring of first grade, while the outcome measure, ORF, was administered in winter and spring. Classroom teachers gave the spelling assessment, DST. School psychology graduate students trained in assessment methodology administered the literacy screeners: PSF, NWF, WIF, and ORF. As part of the larger RTI study, inter-rater reliability data were collected for PSF and NWF. For the PSF measure, average inter-rater reliability was 97.33 in winter and 95.44 in spring, and for NWF, average inter-rater reliability was 97.75 in spring (O'Connor et al., 2010). The English language proficiency measures were both administered at the beginning of the school year. Trained school personnel administered the CELDT to all ELLs, and trained graduate students gave the PPVT.

Statistical analyses. To address the research questions posed in this study, several different statistical analyses were run. For the first research question, Pearson correlation coefficients were calculated to determine the concurrent relations between each of the predictor variables (PSF, DST, NWF, and WIF) at each time point and the outcome measure (ORF) in the spring. A series of multiple regression analyses were conducted to determine how much variance in the outcome measure was explained by each of the predictor variables, and ROC analyses were run to determine appropriate cut scores and examine the predictive accuracy of each early literacy screener. These analyses are discussed in more detail below.

Regression. Hierarchical regression analyses were performed to determine the amount of variance in ORF scores explained by each of the predictor variables (PSF, DST, NWF, and WIF). With hierarchical regression, the predictor variables are entered into the regression analysis in a pre-established order. This controls for the effect of the earlier predictors and allows the researcher to determine how much additional variance (ΔR^2) is explained by each additional predictor (Pedhazur, 1997). The predictors were entered into the regression model based on the order in which the related skills typically develop (i.e., PSF, followed by DST, NWF, and WIF). All the predictor variables were entered to determine how much variance in the outcome was explained by the early literacy assessments.

Further analysis was done to determine if English proficiency, as measured by the CELDT or the PPVT, explained additional variance above and beyond that explained by the early literacy measures. For the purposes of data analysis, the four different CELDT levels (Beginning, Early Intermediate, Intermediate, and Early Advanced) were collapsed into two groups of comparable size: Beginning/Early Intermediate (B/EI, $n = 55$), and Intermediate/Early Advanced (I/EA; $n = 51$). Two dummy variables were created with the NESs serving as the reference group. The subsequent change in the R^2 value (ΔR^2) was examined to determine whether CELDT scores explained additional variance. Similarly, the analysis was run with the early literacy variables entered followed by scores on the PPVT. Regression coefficients (β and *B*), the amount of variance associated with the addition of each predictor variable (R^2 and ΔR^2), and the related *F*-values were reported.

For research question 3, regression analyses were run with each of the predictor variables individually to determine which single predictor explained the most variance. The winter measure that was the strongest predictor was then used as the outcome measure for additional regression analyses with the fall screening assessments as predictors.

Receiver Operating Characteristic (ROC). To answer the last research question, ROC analysis was used to examine classification accuracy. Specifically, ROC curves were generated from each screening measure (PSF, DST, NWF, and WIF) with “at-risk” status on ORF (a score at or below the 25th percentile in spring of first grade) as the outcome being predicted. A ROC curve is a graph of sensitivity (y-axis) by 1 - specificity (x-axis). Once a ROC curve is generated, the area under the curve (AUC) can be examined. The AUC represents the assessment’s ability to discriminate between at-risk and not at-risk students. The closer an AUC gets to 1, the higher its diagnostic accuracy, with a value of 1 indicating that the screening measure differentiates perfectly, and a value of .50 indicating that the screening measure does not predict any better than chance. Swets (1988) stated that AUC results of greater than or equal to .90 are considered to provide good discrimination, values between .70 and .90 are considered useful, and values below .70 are poor.

Cut scores were selected on each measure that gave a sensitivity (true positive rate) as close to .90 as possible. The corresponding specificity (true negative rate) was reported and compared across measures. Then, the data were examined to determine how many cases were predicted accurately. The number of true positives (TP; students predicted to fail who indeed failed), true negatives (TN; students predicted to pass who indeed passed), false positives (FP; students predicted to fail who actually passed), and false negatives (FN; students predicted to pass who actually failed) were calculated. These numbers were used to determine the positive predictive power [TP/(TP + FP)], negative predictive power [TN/(TN + FN)], and hit rate [(TP + TN)/(TP + TN + FP + FN)] of each measure (see Table 2).

To answer question 4, part B, student results were grouped by CELDT proficiency level. ROC analyses were run again for each measure. These analyses were used to determine how AUC, cut scores, and the resulting predictive accuracy vary for students with differing English proficiency levels.

Chapter 4: Results

Descriptive Statistics

Descriptive statistics were run for the entire sample, and Table 3 presents the ranges, means, and standard deviations. In the fall of first grade, students correctly provided an average of 38.24 phonemes (*SD* = 18.42) on PSF, 39.61 phonemes (*SD* = 7.14) on DST, 29.95 sounds (*SD* = 16.67) on NWF, and 12.98 words (*SD* = 12.28) on WIF. As expected, mean scores increased from fall to winter with students correctly providing an average of 49.04 phonemes (*SD* = 13.78) on PSF, 50.61 phonemes (*SD* = 6.51) on DST, 46.09 sounds (*SD* = 19.98) on NWF, and 33.34 words (*SD* = 20.30) on WIF. In spring, students correctly provided an average of 60.96 phonemes (*SD* = 6.77) on DST, 52.90 sounds (*SD* = 24.97) on NWF, and 42.40 words (*SD* = 21.21) on WIF. They correctly read an average of 44.96 words (*SD* = 23.61) on ORF. PPVT standard scores ranged from 45 to 120 with an average of 85.71 (*SD* = 12.13).

Disaggregated by English language proficiency level. Table 3 also includes the descriptive statistics disaggregated by English language proficiency level. For students in the B/EI English language proficiency group, the average scores were consistently below the average of the entire sample. In the fall of first grade, students correctly provided an average of 31.32 phonemes (*SD* = 20.21) on PSF, 37.62 phonemes (*SD* = 8.04) on DST, 25.71 sounds (*SD* = 17.38) on NWF, and 9.07 words (*SD* = 9.53) on WIF. In winter, students correctly provided an average of 48.17 phonemes (*SD* = 12.17) on PSF, 49.62 phonemes (*SD* = 7.24) on DST, 40.73 sounds (*SD* = 16.61) on NWF, and 28.82 words (*SD* = 18.57) on WIF. Scores increased further in spring, with students correctly providing an average of 59.71 phonemes (*SD* = 8.53) on DST, 48.53 sounds (*SD* = 25.14) on NWF, and 39.20 words (*SD* = 20.15) on WIF. Students in this group correctly read an average of 39.85 words (*SD* = 22.31) on ORF, and their PPVT standard scores ranged from 45 to 102 with a mean of 77.18 (*SD* = 11.62).

For students in the I/EA English language proficiency group, mean scores were higher than those of the entire sample, except for PPVT in the fall (85.16 vs. 85.71) and PSF in winter (47.42 vs. 49.04). Specifically, students in the I/EA group correctly provided an average of 41.84 phonemes (*SD* = 16.90) on PSF, 41.39 phonemes (*SD* = 5.23) on DST, 32.18 sounds (*SD* = 16.05) on NWF, and 16.18 words (*SD* = 13.33) on WIF. In winter, students correctly provided an average of 47.43 phonemes (*SD* = 13.08) on PSF, 50.69 phonemes (*SD* = 6.45) on DST, 51.02 sounds (*SD* = 26.51) on NWF, and 38.63 words (*SD* = 21.16) on WIF. Scores increased further in spring, with students correctly providing an average of 62.37 phonemes (*SD* = 4.88) on DST, 55.20 sounds (*SD* = 24.40) on NWF, and 47.82 words (*SD* = 20.12) on WIF. Students in this group correctly read an average of 51.75 words (*SD* = 22.93) on ORF, and their PPVT standard scores ranged from 50 to 111 with a mean of 85.16 (*SD* = 10.39).

Research Question 1: Correlations

Pearson correlation coefficients were calculated to answer the first research question: What are the relations between the early literacy screening measures (PSF, DST, NWF, and WIF) collected in fall, winter, and spring of first grade and the outcome variable, spring ORF? According to Cohen (1992), correlations up to .29 are considered small, correlations between .30 to .49 are considered medium, and correlations at or above .50 are considered large. All correlations were positive, and almost all were significant (the correlation between WIF1 and PSF2 was not, *r* = .09; *p* = .22), though they varied in size from small (*r* = .14) to large (*r* = .91; see Table 4).

Of the literacy screeners administered in the fall, WIF1 was the most strongly correlated with spring ORF (*r* = .72; *p* < .01). There were also large correlations between the fall predictors, DST1 and NWF1, and the outcome variable, ORF3 (*r* = .56; *p* < .01; and *r* = .58; *p* < .01, respectively). Though it was still significant, the correlation between PSF1 and ORF3 was small (*r* = .28; *p* < .01). Similarly, of the winter literacy screeners, WIF2 was the one most strongly correlated with spring ORF (*r* = .91; *p* < .01). Both DST2 and NWF2 were strongly correlated with the outcome variable, ORF3 (*r* = .60; *p* < .01; and *r* = .67; *p* < .01, respectively), as well. Once again, the correlation between PSF2 and ORF3 was small though significant (*r* = .16; *p* < .05).

Examining the concurrent relations among fall screening measures showed that all the measures were significantly correlated with each other. PSF1 was moderately correlated with DST1 (*r* = .39; *p* < .01), NWF1 (*r* = .45; *p* < .01), and WIF1 (*r* = .35; *p* < .01). Similarly, DST1 was moderately correlated with NWF1 (*r* = .48; *p* < .01), and WIF1 (*r* = .46; *p* < .01), while NWF1 was strongly correlated with WIF1 (*r* = .66; *p* < .01). A similar pattern was found when analyzing relations between the winter screeners. Once again, all correlations were significant. However, in this case, PSF2 was only weakly correlated with DST2 (*r* = .26; *p* < .01), NWF2 (*r* = .25; *p* < .01), and WIF2 (*r* = .14; *p* < .05). DST2 was moderately correlated with NWF2 (*r* = .47; *p* < .01) but strongly correlated with WIF2 (*r* = .60; *p* < .01), and NWF2 was strongly correlated with WIF2 (*r*

= .66; *p* < .01). Among the spring screeners, DST3 was moderately correlated with NWF3 (*r* = .42; *p* < .01) but strongly correlated with WIF3 (*r* = .62; *p* < .01). The largest correlation was between NWF3 and WIF3 (*r* = .65; *p* < .01).

Students' receptive vocabulary in English, as measured by the PPVT, was positively and significantly correlated with all the literacy screeners and the outcome measure, though none of the correlations were large (*r* = .15 to .32). The strongest correlations were with DST1 and DST2 (*r* = .32; *p* < .01 in both cases). Small, but significant, correlations were found between PPVT and DST3 (*r* = .22); PSF1 and PSF2 (*r* = .23 for both); NWF1, NWF2, and NWF3 (*r* = .15-.19); and WIF1, WIF2, and WIF3 (*r* = .19-.22). There was also a small but significant relation between PPVT and the outcome variable, ORF3 (*r* = .20; *p* < .01).

RQ1a: Disaggregated correlations. Correlations were run disaggregated by English language proficiency group to answer the second part of research question 1: To what extent do the relations between variables differ for students with different levels of English language proficiency, as measured by CELDT scores? Results are presented in Table 5. Subsequently, the correlations were examined to determine the strength of the relations between the literacy screeners and the outcome measure and how those relations compare across groups.

For students in the B/EI group, WIF1 was the fall screener that was most strongly correlated with spring ORF (*r* = .76; *p* < .01) just as it was for the total sample. There were also large correlations between the fall predictors, DST1 and NWF1, and the outcome variable, ORF3 (*r* = .57 and *p* < .01 for both). However, the correlation between

PSF1 and ORF3 was not significant. Similarly, of the literacy screeners administered in the winter, WIF2 was the one most strongly correlated with spring ORF (*r* = .93; *p* < .01). Both DST2 and NWF2 were strongly correlated with the outcome variable, ORF3 (*r* = .60; *p* < .01; and *r* = .72; *p* < .01, respectively), as well. Once again, the correlation between PSF2 and ORF3 was not significant. The general strength and direction of the correlations were the same as those for the total sample. However, for the total sample, all correlations between the screeners and the outcome variable were significant.

Similarly, for students in the I/EA group, WIF1 was the fall screener that was most strongly correlated with spring ORF (*r* = .70; *p* < .01) just as it was for both the B/EI group and the total sample. There were also large correlations between the fall predictors, DST1 and NWF1, and the outcome variable, ORF3 (*r* = .57; *p* < .01; and *r* = .64; *p* < .01, respectively). As for the B/EI sample, the correlation between PSF1 and ORF3 was not significant. Similarly, of the literacy screeners administered in the winter, WIF2 was the one most strongly correlated with spring ORF (*r* = .90; *p* < .01). Both DST2 and NWF2 were strongly correlated with the outcome variable, ORF3 (*r* = .59; *p* < .01; and *r* = .73; *p* < .01, respectively), as well. Once again, the correlation between PSF2 and ORF3 was not significant. The general strength and direction of the correlations were the same as those for both the B/EI group and the total sample. However, both the B/EI and I/EA group differed from the total sample in that the correlations between PSF and ORF3 were not significant.

For the B/EI group, most of the fall screeners were significantly correlated with each other. PSF1 was not significantly correlated with WIF1, but it was moderately and significantly correlated with both DST1 (*r* = .44; *p* < .01) and NWF1 (*r* = .44; *p* < .01). DST1 was moderately correlated with NWF1 and WIF1 (*r* = .49; *p* < .01 for both), and NWF1 was strongly correlated with WIF1 (*r* = .73; *p* < .01). Similarly, most of the winter screeners were significantly correlated with each other. In this case, PSF2 was significantly correlated with DST2 (*r* = .39; *p* < .01), but not with NWF2 or WIF2. DST2 was moderately correlated with NWF2 (*r* = .49; *p* < .01) but strongly correlated with WIF2 (*r* = .60; *p* < .01), and NWF2 was strongly correlated with WIF2 (*r* = .69; *p* < .01). Among the spring screeners, DST3 was moderately correlated with NWF3 (*r* = .48; *p* < .01) but strongly correlated with WIF3 (*r* = .67; *p* < .01). The strongest correlation was between NWF3 and WIF3 (*r* = .73; *p* < .01).

Most of the concurrent correlations for the B/EI group were similar to those for the total sample in their general strength and direction. However, for the total sample, all concurrent correlations were significant while that was not the case for the B/EI group (as described above). In comparing the correlations found for the B/EI group to those found for the total sample, the largest differences were in the relations between PSF2 and DST2 (*r* = .39; *p* < .01 vs. *r* = .26; *p* < .01) and PSF2 to NWF2 (*r* = .10; *p* > .05 vs. *r* = .25; *p* < .01). For this group of students, their receptive vocabulary in English, as measured by the PPVT, was moderately and significantly correlated with PSF1 and PSF2 (*r* = .33; *p* < .05; *r* = .30; *p* < .05, respectively) and DST1, DST2, and DST3 (*r* = .30-.37) but not with NWF, WIF, or ORF.

For the I/EA group, all the fall screeners were significantly correlated with each other except for PSF1 and DST1 (*r* = .18; *p* > .05). PSF1 was moderately and significantly correlated with both NWF1 (r = .43; p < .01) and WIF1 (r = .39; p < .01). DST1 was moderately correlated with NWF1 (r = .46; p < .01) and WIF1 (r = .48; p < .01), and NWF1 was strongly correlated with WIF1 (r = .59; p < .01). Similarly, most of the winter screeners were significantly correlated with each other. In this case, PSF2 was significantly correlated with DST2 (r = .30; p < .05), but not with NWF2 (r = .28; p > .05) or WIF2 (r = .19; p > .05). DST2 was strongly correlated with NWF2 (r = .50; p < .01) and with WIF2 (r = .58; p < .01), and NWF2 was strongly correlated with WIF2 (r = .66; p < .01). Among the spring screeners, DST3 was moderately correlated with NWF3 (r = .37; p < .01) but strongly correlated with WIF3 (r = .58; p < .01). The strongest correlation was between NWF3 and WIF3 (r = .68; p < .01).

Most of the concurrent correlations for the I/EA group were also similar to those for the total sample in their general strength and direction. However, for the total sample, all concurrent correlations were significant while that was not the case for the I/EA group (as described above). Comparing the I/EA group with the B/EI group and the total sample revealed a difference in the relations between PSF1 and DST1. While both the B/EI group and the total sample showed a moderate correlation between PST1 and DST1 (r = .44; p < .01, and r = .39; p < .01, respectively), the correlation was weak and non-significant for the I/EA group.

For students in the I/EA group, their receptive vocabulary in English, as measured by the PPVT, was significantly correlated with only one of the fall screeners (DST2; r = .44; p < .01). However, it was significantly correlated with all the winter variables (r = .31-.34; p < .05), and with both DST3 (r = .37; p < .01) and WIF3 (r = .32; p < .05).

Though PPVT was not significantly correlated with ORF3 for the B/EI group, it was moderately correlated for the I/EA group (r = .35; p < .05).

Research Question 2: Regression Analyses and Variance Explained

This section addresses the second research question: How much variance in reading outcomes (spring ORF) is explained by the early literacy predictor variables (PSF, DST, NWF, and WIF) administered in fall and winter? A series of hierarchical regression analyses were conducted with variables entered in the order that the assessed skills were expected to develop, i.e., PSF, DST, NWF, and WIF. According to Cohen (1988), R^2 values of .02, .13, and .26 correspond with small, medium, and large effect sizes, respectively. Results are presented in Tables 6 and 7 and summarized below.

For the model using fall screeners as predictors, PSF1 was entered into the model first and was statistically significant, though it only explained 8.1% of the variance in ORF3. Both DST1 and the quadratic form, DST1^2, were entered next. The quadratic form was included since examination of a scatterplot of DST1 and ORF3 revealed a non-linear relation between them, and comparison of the linear and quadratic models found the quadratic model to be the best fit. Entering these variables into the model resulted in a large and significant increase in R^2 [$F_{(3, 205)} = 48.59$, p <.001, $\Delta R^2 = .30$]. Though PSF1 was initially significant (B = .36, p < .001), entering DST1 and DST1^2 into the model made PSF1 non-significant (B = .08, p > .05). Adding NWF1 in step 3 also led to a significant increase in R^2 [$F_{(4, 204)} = 39.71$, p <.001, ΔR^2= .10], and with the addition of NWF to the model, DST^2 and NWF were the only significant variables. Finally, WIF1 was added in Step 4, which significantly increased R^2 [$F_{(5, 203)} = 62.49$, p <.001, ΔR^2 =

.12] and made all the other predictors insignificant. Altogether, the final model, ORF3' = 12.659 - 0.08 PSF1 - 0.35 DST1 + 0.02 DST1^2 + 0.18 NWF1 + 0.96 WIF1, explained 60% of the variance in ORF3. Standardized beta values (β) for PSF, DST, DST^2, NWF, and WIF were -0.07, -0.11, 0.42, 0.12, and 0.50, respectively.

To determine how much variance in spring ORF was explained by the winter variables, a second hierarchical regression model was run following the same sequence of steps. PSF2 was entered into the model first and, once again, it was statistically significant, though it only explained a small percentage of the variance in ORF3 (2.7%). DST2 and the quadratic form, DST2^2, were entered in step 2 which resulted in a large and significant increase in R_2 [$F_{(3, 205)} = 79.96$, p <.001, $\Delta R^2 = .43$]. As with the fall model, entering DST2 and DST2^2 into the model made PSF1 non-significant (B = .01, p > .05). Adding NWF2 in step 3 led to another significant increase in R^2 [$F_{(4, 204)} = 76.46$, p <.001, $\Delta R^2 = .15$], and in this case, DST2, DST2^2 and NWF2 were all significant (B = -3.64, p <.01; B = .06, p <.001; and B = .53, p <.001, respectively). Finally, WIF2 was added in Step 4, which significantly increased R^2 [$F_{(5, 203)} = 315.19$, p <.001, $\Delta R^2 = .24$] and made the DST variables insignificant. Altogether, the final model, ORF3' = 10.43 + 0.01 PSF2 - 0.51 DST21 + 0.01 DST2^2 + 0.13 NWF2 + 0.91 WIF2, explained 84% of the variance in ORF3. Standardized beta values (β) for PSF, DST, DST^2, NWF, and WIF were 0.01, -0.14, 0.23, 0.11, and 0.78, respectively.

RQ2a: Additional variance explained by CELDT. An additional step was added to the hierarchical regression model to answer part A of research question 2: To what extent does English language proficiency, as measured by CELDT scores, explain additional variance in outcomes after controlling for student performance on early literacy measures? As described previously, PSF, DST, NWF, and WIF were entered step-by-step into the fall and winter models. Dummy variables were used to indicate membership in the B/EI or I/EA groups. NESs served as the reference group.

For the fall model, the screening measures accounted for 60% of the variance in ORF3. After controlling for the contribution of those measures, the two language proficiency variables entered together in step 5 did not significantly increase the explained variance in ORF3 scores. Similarly, neither of the dummy variables were statistically significant, indicating that, on average, the ELL students' scores on spring ORF were not significantly different than those of the NESs. Nevertheless, the final model was statistically significant and accounted for 61% of the variance in spring ORF, $F_{(7, 201)} = 43.677$, $p < .001$.

For the winter model, similar results were found. The screening measures already accounted for 84% of the variance in ORF3. After controlling for the contribution of those measures, the two language proficiency variables entered together in step 5 did not significantly increase the explained variance in ORF3 score nor were the dummy variables statistically significant. The final model was statistically significant and accounted for 84% of the variance in spring ORF, $F_{(7, 201)} = 154.18$, $p < .001$.

RQ2b: Additional variance explained by PPVT. The same hierarchical regression model was run with PPVT scores substituted for CELDT groups in step 5. This provided the answer to part B of research question 2: To what extent does one aspect of English language proficiency, receptive vocabulary as measured by PPVT scores, explain additional variance in outcomes after controlling for student performance on early literacy measures? However, as with the CELDT, the PPVT did not significantly increase explained variance, nor was it statistically significant in either the fall or winter models (see Tables 5 and 6).

Research Question 3: Single Best Predictor Winter to Spring

This section addresses research question 3: Considering screeners from within a literacy development framework, which winter early literacy measure explains the most variance in spring ORF? The hierarchical regressions run previously found that NWF2 and WIF2 were the only significant variables in the final model for winter (β = .11, β = .78, respectively; p < .01). Also, adding WIF2 to a model that already included NWF2 explained additional variance. When entered into a regression analysis as the single predictor, WIF2 explained 83.1% of the variance in ORF3, $F_{(1, 207)}$ = 1020.33, p < .001.

Research question 3a: Single best predictor fall to winter. Identifying winter WIF as the strongest predictor of spring ORF makes it possible to answer part A of research question 3: Which of the fall early literacy measures accounts for the most variance in that winter predictor? For this regression analysis, the fall early literacy measures were once again entered into the model in order (see Table 8). PSF1 was entered into the model first and was statistically significant, though it only explained 8.7% of the variance in WIF2. Both DST1 and the quadratic form, DST1^2, were entered next. The quadratic form was included since examination of a scatterplot of DST1 and WIF2 revealed a non-linear relation between them, and comparison of the linear and quadratic models found the quadratic model to be the best fit. Entering these variables into the model resulted in a large and significant increase in R^2 [$F_{(3, 205)} = 50.67$, p <.001, $\Delta R^2 = .30$]. Though PSF1 was initially significant, entering DST1 and $DST1^2$ into the model made PSF1 non-significant. Adding NWF1 in step 3 also led to a significant increase in R^2 [$F_{(4, 204)} = 47.07$, p <.001, $\Delta R^2 = .12$]. Finally, WIF2 was added in Step 4, which significantly increased R^2 [F ($_{5, 203)} = 52.58$, p <.001, $\Delta R^2 = .10$] and made DST1 insignificant. This final model explained 61% of the variance in WIF2 and yielded the following regression equation: WIF2' = 6.48 - 0.07 PSF1 – 0.61 DST1 + 0.02 $DST1^2$ + 0.21 NWF1 + 0.75 WIF1. Of the fall screeners, PSF and the linear form of DST were the only variables that were not significant at the .05 level. Standardized beta values (β) for PSF1, DST1, $DST1^2$, NWF1 and WIF1 were -0.06, -0.21, 0.53, 0.17 and 0.45, respectively.

Research Question 4: ROC Analyses and Predictive Accuracy

This section uses ROC analyses to answer research question 4: What is the accuracy (area under the curve, sensitivity, and specificity) of PSF, DST, NWF, and WIF in predicting scores below the 25th percentile on spring ORF? There are multiple steps involved in answering this question. With the current sample, the first step was to determine that a score of 28 on ORF3 was equivalent to the 25th percentile. Students who scored below this were considered below expectations while those who scored a 28 or higher were considered to have met or exceeded expectations on spring ORF.

Using that cutoff, ROC curves were generated, and AUC values were evaluated to examine the accuracy of each of the literacy screeners in predicting spring ORF. AUC results greater than or equal to .90 are considered to provide good discrimination, values between .70 and .90 are considered useful, and values below .70 are poor (Swets, 1988). By these criteria, the AUC value for fall PSF (AUC = .69; 95% CI [.61, .77]) was poor, but the values for fall DST (AUC = .84; [.78, .91], NWF (AUC = .84; [.77, .90], and WIF (AUC = .87; [.82, .92]), all fell within the useful range. For the winter screeners, the AUC value for PSF was still poor (AUC = .63; 95% CI [.54, .71]), values for DST (AUC = .87; [.81, .92] and NWF (AUC = .81; [.75, .88] were useful, and the value for WIF (AUC = .97; [.96, .99] was quite good.

In addition, the range of scores for each measure was evaluated to determine the optimal sensitivity and specificity. Specifically, the Youden index (*J*) was calculated across a range of scores to determine the point where sensitivity and specificity are maximized (*J* = max[*sn* + *sp*]). Table 9 lists the maximal sensitivity and specificity for each measure along with the cut score associated with those values. Of the fall screeners, PSF had the lowest sensitivity and specificity (*sn* = .61; *sp* = .75). DST and NWF were similar to each other with sensitivity that was slightly lower than specificity. (*sn* = .71-.75; *sp* = .81-.85). WIF had the highest sensitivity but a lower specificity value (*sn* = .82; *sp* = .76). Of the winter screeners, PSF once again had the lowest sensitivity (*sn* = .45; *sp* = .73). DST, NWF, and WIF all had fairly high sensitivity (.88-.92) but winter WIF stood out with both sensitivity and specificity above 90% (*sn* = .92; *sp* = .92).

Research question 4a: Cut scores. However, using a formula to calculate the maximal sensitivity and specificity values cannot address whether those values are appropriate for a specific use. For this reason, the analysis was taken a step further to answer part A of research question 4: Which cut scores provide the best combination of sensitivity and specificity for screening purposes? For screening measures, Rathvon (2004) argues that sensitivity and specificity should both be at least 75-80% to match reliability standards (Salvia et al., 2010), but other researchers argue for a higher standard. Jenkins et al. (2007) recommend a minimum standard of 90% for sensitivity to prevent missing truly at-risk students, and Compton et al. (2010) recommend a minimum standard of 80% for specificity so that schools' resources are not spread too thin. By these standards, only DST2 and WIF2 demonstrated sufficient sensitivity.

For this reason, the next step was selecting cut scores on each measure to yield a value for sensitivity as close to .90 as possible. The corresponding specificity was reported, and the data were examined to determine how many cases were predicted accurately. Table 10 reports the cut score selected for each measure along with the corresponding sensitivity, specificity, number of true positives, true negatives, false positives, and false negatives. The positive predictive power, negative predictive power, and hit rate of each measure were also calculated and included in the table.

Among the 209 students in this sample, 51 (24% of the sample) scored below expectations on spring ORF while 158 met or exceeded expectations. For fall PSF, NWF, and WIF, selecting the cut score that corresponded with 90% sensitivity meant that the screener correctly identified as at-risk 46 students (true positives; TP) out of the 51 who scored below expectations on spring ORF while missing 5 (false negatives; FN). Fall DST was slightly different and correctly identified 45 at-risk students while missing 6.

Of the 158 students who met or exceeded expectations on spring ORF, fall PSF correctly predicted that 57 were not at-risk (true negatives; TN), but 101 were incorrectly identified as at-risk (false positives; FP) which means this measure had poor sensitivity (36%). Predictive power is another way to calculate a screening measure's accuracy in predicting the outcome (see Table 2). Out of 209 students, the PSF cut score used predicted that 147 were at-risk (70% of the sample) when only 46 truly were, which resulted in a positive predictive power of only 31%. Conversely, 62 students were identified as not at-risk, and this prediction was accurate for 57 of them resulting in a negative predictive power of 92%. Hit rate measures the overall accuracy and, for fall PSF, it was only 49%.

Fall DST correctly predicted 94 of the 158 students who met or exceeded expectations on spring ORF (TN). However, 64 students were incorrectly identified as at-risk (FP) which means this measure also had low sensitivity (60%). The cut score used predicted that 109 students were at-risk (52% of the sample) when only 45 truly were, which resulted in a positive predictive power of only 41%. Conversely, 100 students were identified as not at-risk, and this prediction was accurate for 94 of them resulting in a negative predictive power of 94%. Hit rate for fall DST was 67%.

Fall NWF's accuracy was slightly below that of DST. NWF correctly predicted 72 of the 158 students who met or exceeded expectations on spring ORF (TN). However, 86 students were incorrectly identified as at-risk (FP) which resulted in sensitivity value of 46%. The cut score used predicted that 132 students were at-risk (63% of the sample) when only 46 truly were, which resulted in a positive predictive power of 35%. Conversely, 77 students were identified as not at-risk, and this prediction was accurate for 72 of them resulting in a negative predictive power of 94%. Hit rate for fall NWF was 56%.

Fall WIF was the best predictor, though not by much. WIF correctly predicted 103 of the 158 students who met or exceeded expectations on spring ORF (TN). However, 55 students were incorrectly identified as at-risk (FP) which resulted in sensitivity value of 65%. The cut score used predicted that 101 students were at-risk (48% of the sample) when only 46 truly were which resulted in a positive predictive power of 46%. Conversely, 108 students were identified as not at-risk, and this prediction was accurate for 103 of them resulting in a negative predictive power of 95%. Hit rate for fall WIF was 71%.

Of the winter screeners, PSF and NWF were both less accurate than they were in the fall, with hit rates of 41% and 48% respectively. However, DST2 and WIF2 both had improved accuracy. Winter DST correctly predicted 106 of the 158 students who met or exceeded expectations on spring ORF (TN). Fifty-two students were incorrectly identified as at-risk (FP) which means this measure still had low sensitivity (67%). The cut score used predicted that 98 students were at-risk (47% of the sample) when only 46 truly were which resulted in a positive predictive power of 47%. The DST2 cut score identified 111 students as not at-risk, and this prediction was accurate for 106 of them resulting in a negative predictive power of 94%. Hit rate for winter DST was 73%.

Winter WIF was the most accurate predictor. This screener correctly predicted 146 of the 158 students who met or exceeded expectations on spring ORF (TN). Only 12 students were incorrectly identified as at-risk (FP) which resulted in sensitivity value of

92%. The cut score used predicted that 59 students were at-risk (28% of the sample) and 47 of those truly were, which resulted in a positive predictive power of 80%. Of the 150 students identified by WIF2 as not at-risk, the prediction was accurate for 146 of them resulting in a negative predictive power of 97% and an overall hit rate of 92%.

Research question 4b: Predictive accuracy for different ELL groups. Student results were grouped by CELDT proficiency level and analyzed to answer research question 4, part B: To what extent is there a difference in predictive accuracy for students with differing English proficiency levels as measured by the CELDT? ROC analyses were run again for each measure to determine how AUC, cut scores, and the resulting predictive accuracy varied for students with differing English proficiency levels.

For the ELLs in both the B/EI and I/EA groups, the AUC for each of the fall measures was similar to those calculated for the sample as a whole (see Table 11). For both groups, the AUC for PSF1 was too low to be considered useful (.60-.65), while DST1, NWF1, and WIF1 all fell within the useful range (.77-.90). For the B/EI group, WIF1 had the highest AUC of the fall screeners at .85 (95% CI [.74, .95]), but for the I/EA group, NWF was the highest (AUC = .90; 95% CI [.77, 1.00]).

For fall PSF, overall accuracy as measured by the hit rate was the same for both ELL groups (.45) and similar to that found for the total sample (.49). DST1's accuracy was lower for the B/EI and I/EA groups than for the total sample (.60 and .55, respectively, vs. .67). For the B/EI group, the hit rate for fall NWF was comparable to that of the total sample (.53 vs. .56), but for the I/EA group, NWF1 had a much higher hit rate (.94) and was the most accurate measure. The hit rate for fall WIF was .71 for the total sample but ranged from .62 for the B/EI group to .76 for the I/EA group.

With cut scores chosen to keep sensitivity as close to 90% as possible, specificity ranged from .28 to .50 across measures for the B/EI group. For the I/EA group, PSF1 and DST1 both had low specificity (.37 and .49, respectively), while WIF1 had acceptable specificity (.74), and NWF1's was excellent (.95).

Similar results were found for the winter screeners (see Table 12). Once again, for the ELLs in both the B/EI and I/EA groups, the AUC for each of the winter measures was comparable to that calculated for the sample as a whole. For both groups, the AUC for PSF2 was too low to be considered useful (.55-.61), while DST2 and NWF2 fell into the useful range (.81-.83). As with the total sample, winter WIF stood out as the strongest predictor with AUCs in the good range (.96-.99).

Winter PSF's overall accuracy as measured by the hit rate was low for the B/EI group (.42) and even lower for the I/EA group (.27). DST2's accuracy was lower for the B/EI and I/EA groups than for the total sample (.65 and .61, respectively, vs. .73), but NWF2's accuracy was higher for the ELL groups (.65 and .78 respectively; vs. .48). For both groups, the hit rate for winter WIF was the highest of the winter screeners, ranging from .90 for the I/EA group to .95 for the B/EI group.

For the B/EI group, specificity ranged from .23 to .55 on winter PSF, DST, and NWF, but was much higher for WIF (.95). For the I/EA group, PSF2 had extremely low specificity (.16), while DST2's was .56. The specificity for NWF2 was acceptable at .77, while WIF2's was the highest at .91.

Research question 4c: Cut scores for ELLs. This section addresses part C of research question 4: Is there a practical difference in cut scores obtained for students with differing English proficiency levels as measured by the CELDT? For the fall screeners, cut scores were quite similar across groups (see Table 13). On fall PSF, the identified cut scores were 49, 49, and 50 for the total sample, the B/EI group, and the I/EA group, respectively. Similarly, fall DST cut scores were almost equivalent across groups (42, 42, and 43, respectively) as were the cut scores for WIF (10, 9, and 10, respectively). However, the cut score for fall NWF was much lower for the I/EA group (16) than for either the total sample (34) or the B/EI group (30).

If a cut score of 30 had been used for the I/EA group, the sensitivity would have remained unchanged (.88), but the specificity would have dropped to .58 and the overall hit rate to .63. Practically speaking, using the higher cut score would have meant that 18 students would have been falsely identified as at-risk compared with only 2 when using the lower cut score.

For the winter screeners, however, no major differences were identified. On winter PSF, the cut scores across groups ranged from 60 to 62. For winter DST, they varied from 52 to 53. A larger difference was found for winter NWF with a cut score of 49 for the total sample and 40 for both the B/EI and I/EA groups, and there were small differences across groups for winter WIF with cut scores ranging from 17-21.

Chapter 5: Discussion

The major purposes of this study were twofold. The first goal was to compare early literacy screeners across first grade and identify the one that can best predict Oral Reading Fluency (ORF) scores at the end of first grade. The second goal was to determine whether Spanish-speaking English language learners' (Ss ELLs) level of English language proficiency influences a screener's predictive power and classification accuracy.

Based on the order in which literacy skills develop and on findings from prior screening studies, it was hypothesized that Phoneme Segmentation Fluency (PSF) would be the weakest predictor, Nonsense Word Fluency (NWF) would be stronger, and Word Identification Fluency (WIF) would be the strongest predictor. The measure of invented spelling, the Developmental Spelling Test (DST), was expected to be stronger than PSF since it measures both phonological awareness and orthographic skills. However, evidence from prior research was not sufficient to predict how DST would compare to other screeners.

Some screeners have been researched with Ss ELLs more thoroughly than others. Though there was not enough evidence to predict whether level of English language proficiency would impact classification accuracy for PSF, DST, or WIF, some research has found that ELL status changes the appropriate cut score for NWF (Johnson et al., 2009; Vanderwood et al., 2008).

Overall, findings support previous research indicating that, for first graders, WIF is the best predictor of later reading performance. Results from this study extend the current literature on screening assessments by examining each measure’s ability to predict future reading performance for Ss ELLs at different levels of English language proficiency, and findings indicate that the same cut scores can be used for ELLs at each level as for native English speakers (NESs). This study also adds to the knowledge base on the use of invented spelling as a screening assessment by examining its predictive power and classification accuracy.

Correlations Between Measures

Predictive. Consistent with prior studies (e.g., Clemens et al., 2011; Healy, 2007; Munger & Murray, 2017), DST, NWF, and WIF were all strongly correlated with the outcome measure, ORF3, and the strength of the correlations increased from fall to spring, as expected. PSF’s relations with ORF3 were the only exception to this pattern with weak correlations that decreased in size from fall to winter. However, even this apparent surprise is consistent with expectations if one takes into account the early literacy skills being measured and how they develop.

Prior studies have found small to moderate correlations between PSF and ORF (Clemens et al., 2011; Munger & Blachman, 2013; Munger & Murray, 2017; Vanderwood et al., 2014), and phonological awareness develops well before the ability to read connected text. PA is foundational to developing the alphabetic principle, and scores on measures of segmenting tend to decline once the subsequent skills of decoding and blending are established and students begin to decode patterns in words rather than individual letter sounds. During data collection for this study, the sixth edition of DIBELS was used, but an updated version, known as DIBELS Next, has since been released. In this new version, PSF has been removed from the first-grade winter and spring screening batteries (University of Oregon Center on Teaching & Learning, 2012).

On the other end of the spectrum, WIF was the most strongly correlated with ORF, which is consistent with the findings from other screening studies (e.g., Clemens et al., 2011; Compton et al., 2010; Healy, 2007; Han et al., 2015). Again, this makes sense from a developmental perspective and reflects the similarity of the tasks required in each assessment. Clearly, the ability to quickly read a list of high-frequency words is similar to the ability to quickly read words in connected text.

The correlations between DST and ORF3 and between NWF and ORF3 were similar in size, though NWF's were slightly larger. NWF has been studied extensively and found to have large correlations with ORF3 (e.g., Clemens et al., 2011; Fien et al., 2008; Healy, 2007; Munger & Murray, 2017). However, there is limited research on DST. Munger and Murray (2017) used the Test of Phonological Awareness-Second Edition: PLUS (TOPA-2+) in combination with a modified version of Tangel and Blachman's (1992; 1995) scoring system to generate an invented spelling score for each student. Specifically, the Letter-Sounds subtest was used, and students were asked to spell phonetically regular nonwords. Their invented spelling measure was strongly correlated with ORF3 (r = .53), though the correlation between NWF3 and ORF3 was considerably stronger (r = .89).

Concurrent. Overall, the pattern of concurrent correlations was consistent with prior research. The largest correlations between screeners were between NWF and WIF. The strong correlations observed are consistent with other studies (e.g., Clemens et al., 2011; Healy, 2007). However, though PSF's correlations with the other screeners were significant, they were not strong. This matches prior studies (Johnson et al., 2009; Clemens et al., 2011; Burke et al., 2009, Munger & Blachman, 2013).

While DST's correlations with NWF and WIF were initially small to moderate, they strengthened in the winter and spring. In this study, DST was more strongly correlated with WIF than with NWF at each time point, mostly likely because both DST and WIF test students' knowledge of real words while NWF does not. Due to the relatively limited research on both WIF and DST as a first-grade screener, there are no comparable studies which reported correlations between those two measures. However, when Munger and Murray (2017) used a modified version of the DST in the spring of first grade, they found strong correlations between it and both PSF3 (r = .58) and NWF3 (r = .60). In their study, students were asked to spell phonetically regular nonwords, which explains the increased strength of the correlation between their modified measure and NWF.

Language proficiency. The Peabody Picture Vocabulary Test (PPVT), a measure of receptive vocabulary, was used as a proxy for language proficiency in this analysis. It was positively and significantly correlated with all the screeners and the outcome measure, though the correlations were small. Interestingly, the strongest correlations were with DST1 and DST2 which would indicate that performing well on this measure is more closely linked with vocabulary than it is for the other screeners. On the flip side, it is unsurprising that NWF was the least strongly correlated with vocabulary, as that measure does not use real words like the others do.

When the correlations were disaggregated by CELDT level, the general strength and direction of the correlations were the same as those found for the total sample, regardless of students' English language proficiency level. Though there is limited research with ELLs, this pattern is consistent with what others have found for WIF and NWF (Fien et al., 2008; Healy, 2007; Han et al., 2015; Vanderwood et al., 2008). Of these studies, Han and colleagues (2015) were the only ones who examined ELLs in separate subgroups based on English language proficiency. They found considerable variation in correlations across subgroups. Correlations between WIF2 and ORF3 ranged from a low of .77 for the Early Advanced/Advanced (EA/A) group to a high of .91 for the B/EI group. Along the same lines, correlations between NWF2 and ORF3 were lowest for the EA/A group (r = .31) and highest for the B/EI group (r = .71). Unlike Han and colleagues, the variations in correlations across subgroups were not significant in this study.

The discrepancy between this study's results and those of Han and colleagues is likely due to differences in the ELLs' native language. Han et al.'s study was with Korean-speaking ELLs. While Spanish and English both share root words from Latin, Korean and English are not as closely related. For example, although Korean is an alphabetic language, its orthography has a nonlinear spatial layout, unlike Spanish or English. Such language differences may impact how literacy skills develop and lead to different challenges for ELLs who speak Korean than for those who speak Spanish.

Summary of Regression Analyses

For both the fall and winter hierarchical models, each screener entered (PSF, DST with DST^2, NWF, and WIF) accounted for a significant increase in explained variance for spring ORF. These results are consistent with what would be expected based on the early literacy development process described earlier. However, not all screeners remained significant. In the fall, WIF1 was the only significant predictor once all the predictors were entered into the model, but in the winter, both NWF and WIF were significant predictors. This makes sense given how similar WIF is to the outcome measure, ORF (Zumeta et al., 2012).

Nonetheless, these findings contrast with data from Clemens and colleagues (2011). They used multiple screeners (LNF1, PSF1, NWF1, and WIF1) to predict spring ORF and found that both PSF1 and WIF1 were significant predictors. Only 5% of Clemens and colleagues' sample were ELLs, and though their initial PSF scores were similar to those for this study's sample, their ORF scores at the end of the year were higher (M = 57.96) with a greater standard deviation (40.34) and range (3-165). It is possible that the greater variance in ORF scores made it easier to detect the small influence of PSF in predicting ORF3 or that the inclusion of DST in this study's model reduced the amount of variance uniquely explained by PSF.

Healy (2007) evaluated the ability of NWF1 and WIF1 to predict ORF3. While she found that both screeners were significant, NWF1 only explained 2% of the variance compared to 49% explained by WIF1. Similarly, in the winter, both NWF2 and WIF2 were significant predictors, but WIF2 explained much more variance. The difference between Healy's findings and those in this study may be due to the inclusion of other variables in the model. If data from this study are used to run a regression model with only NWF1 and WIF1 predicting ORF3, then WIF1 explains 52% of the variance, NWF1 explains 2%, and both variables are significant. These results are almost identical to Healy's findings. Including PSF and DST in this study's model reduced the amount of variance uniquely explained by NWF.

Language proficiency. To measure the impact of language proficiency, dummy variables for the CELDT groups were added into the model after the literacy screeners. This resulted in a small, nonsignificant change in R^2. Han and colleagues found very similar results in their 2015 study. They added English proficiency dummy variables to a model that already included PSF2, NWF2, and WIF2. The English proficiency variables only increased the explained variance in ORF3 by 1%, and that change was not significant. Along the same lines, Healy (2007) found that language status (NES vs. ELL) alone was significant but only explained 2% of the variance in ORF3. Once a literacy screener was added to the model, language was no longer a significant predictor.

A similar model was run with PPVT added to the literacy screeners instead of CELDT. PPVT was used because it is a continuous variable, and the CELDT data were categorical. However, the CELDT is a broader measure of language proficiency than PPVT, which only measures receptive vocabulary. Given that the CELDT variable was insignificant, it is unsurprising that PPVT was also nonsignificant, despite the significant correlation between PPVT and ORF3. While vocabulary and fluency are related, the effect of vocabulary upon fluency is likely mediated by the other early literacy skills included in the predictive model**.**

Single best predictor. As discussed previously, WIF2 was easily identified as the strongest single predictor of spring ORF, so an additional regression analysis was run to determine which fall screener best predicted winter WIF. When examining fall variables that could predict winter WIF, the same step-by-step entry was used to create the regression model. As with the ORF models, each screener entered into the winter WIF model was initially significant. However, only NWF1 and WIF1 remained significant once all screeners were entered. Munger and Murray found similar results in their 2017 study. In a hierarchical model predicting WIAT-II Word Reading scores, PSF3 was not significant, NWF3 explained 60% of the variance, and the addition of Invented Spelling (a measure very similar to DST) only increased R-squared by .01 and was non-significant.

Classification Accuracy

Area under the curve. Using Swet's (1988) criteria, winter WIF stood out as the strongest predictor of ORF3 and was the only screener with an area under the curve (AUC) in the range considered to provide good discrimination (above .90). Although not all screening studies calculate or report AUC (e.g., Healy, 2007; Jenkins et al., 2007; Han et al., 2015), Clemens and colleagues (2011) found an AUC of .89 for fall WIF when predicting the 30th percentile on ORF3, which is very similar to the AUC of .87 found for fall WIF in this study.

Though fall WIF and winter WIF both had high AUCs (.87 and .97, respectively), the difference in accuracy between the two illustrates how difficult it can be to predict future performance for first graders as they are learning to read. With sensitivity held as close to .90 as possible, fall WIF misclassified 21% more students (43 out of 209) as at-risk than winter WIF did. Providing intervention services for 26% of first graders (vs. 5% identified in winter) is likely to place a significant burden on a school’s resources.

Interestingly, DST and NWF had very similar AUCs. They were the same in the fall, and DST’s was slightly higher than NWF’s in the winter. Although there is little research on DST as a screener and none that reports AUC, NWF has been extensively researched. This study’s findings are quite similar to those found by other researchers. Johnson et al. (2009) reported an AUC of .82 for fall NWF as a predictor of the 20th percentile on the SAT at the end of first grade, while Clemens and colleagues (2012) reported an AUC of .87 for fall NWF when predicting the 30th percentile on ORF3. In the same study, Clemens and colleagues found that NWF's AUC fell slightly from fall to winter, which matches my findings. However, Catts et al. (2009), in their study predicting third grade ORF, found almost exactly the same value in the winter of first grade as in the fall (.84 vs. .85), and Smolkowski and Cummings (2016) found a slight increase from an AUC of .84 in the fall to .87 in the winter when predicting SAT10 scores at the end of first grade.

Unlike the other screeners, PSF’s AUC was low enough to be considered a poor predictor. This is consistent with other researchers' results, whether predicting ORF3 (e.g., Catts et al., 2009; Clemens et al., 2012) or the SAT at the end of first grade (Johnson et al., 2009). In my study, the AUC value for PSF decreased from fall to winter, which is also consistent with prior research (e.g., Catts et al., 2009; Clemens et al., 2012; Smolkowski & Cummings, 2016). Though this drop in AUC is unusual for other screeners which tend to improve with proximity, it is consistent with PSF's role as a precursor to developing the alphabetic principle and the tendency for scores on PSF to decrease for both struggling readers and skilled ones once they start to move beyond segmenting to blending and more advanced decoding skills.

Language proficiency. When ROC analyses were run by language proficiency group, the results were quite similar to those found for the total sample which would indicate that the classification accuracy of each screener is not influenced by students' level of English language proficiency. The only noticeable difference occurred with the I/EA group. For these students, fall NWF had the highest AUC, rather than WIF. While there were no comparable studies with ELLs that reported AUC, it appears that decoding skills are more closely related to reading success for this group of students, Ss ELLs who have already developed some English language skills.

Cut scores.

WIF. Just as WIF was the strongest predictor when evaluated by AUC values, it was also the measure with the highest sensitivity and specificity when cut scores were selected to result in a sensitivity as close to .90 as possible. The results for fall WIF were quite similar to Clemens and colleagues' (2011) findings. When predicting below the 30th percentile on ORF3, they used a cut score of 12 and found sensitivity of .90 with specificity of .61.

In the winter, WIF2's sensitivity and specificity were both above .90, which is comparable with Han and colleagues' (2015) findings. In their study, winter WIF had a sensitivity of 1.00 and specificity of .92. However, their cut score of 26 was quite a bit higher, probably because they had a more skilled sample and a lower base rate.

Conversely, Healy (2007) also found high sensitivity and specificity (.87 and .86, respectively). However, she used a much lower cut score (WIF2 = 4) when predicting At-Risk status on ORF3. Though Healy's sample started the year with mean scores similar to those of students in this study (e.g., a mean of 13.70 on WIF1 vs. 12.98), they ended the year with much higher means (WIF3 *M* = 58.11 vs. 42.40; ORF3 *M* = 64.02 vs. 44.96). This rate of growth is likely due to the instruction and intervention services that students received and may explain why a lower cut score was used. Only the students who were struggling the most at the beginning of the year were unable to hit the ORF target at the end of the year.

NWF. Though the AUC for NWF fell into the useful range, it did not meet the recommended criteria for a screener. When cut scores were chosen to maximize sensitivity, the corresponding specificity values were too low. Though the cut scores obtained in this study were considerably higher than the DIBELS At-Risk cut scores (13 for fall and 30 for winter), both the cut scores and the specificity values were similar to those found by other researchers.

For fall NWF, Johnson et al. (2009) used a cut score of 38, which resulted in a specificity of .50 when predicting SAT scores at the end of first grade. However, Clemens et al. (2011) used a much lower cut score (23) to get a specificity of .59 when predicting scores below the 30th percentile on ORF3. Comparing the sample's mean scores with those found in this study reveals that they grew more quickly. As with

Healy's WIF findings, this rate of growth was most likely due to the instruction and intervention services that students received as part of Clemens and colleagues' response to intervention (RTI) study. The higher rate of growth may have helped borderline students to hit the ORF target by the end of the year.

For winter NWF, Han and colleagues (2015) used a cut score of 50 and found a specificity of .68 when predicting ORF3. However, Goffreda et al. (2009) used a cut score of only 24 to get a sensitivity of .95 with a specificity of .55 when predicting proficiency on the reading subscale of the TerraNova assessment at the end of second grade. In this case, the difference in cut scores may be a reflection of the length of time between the screener and the outcome measure. Though reaching proficiency on the TerraNova subscale is more challenging than the ORF3 target in this study, the additional year of growth likely increased the range of skill levels and allowed the research team to differentiate using a lower cut score.

DST. In this study, both fall and winter DST were similar to NWF in the AUC values found and had higher specificity than NWF did when cut scores were chosen to correspond with sensitivity of approximately .90. Unfortunately, the results for this measure are difficult to evaluate as no comparable studies reported the results of a ROC analysis. The strength of the measure likely comes from being a broader measure that NWF, so it captures some variation from phonological awareness and orthographic skill.

PSF. Since the AUC values for this measure were the lowest, it was no surprise to find that PSF also had the lowest specificity values when cut scores were selected to obtain sensitivity as close to .90 as possible. The poor performance of PSF at accurately classifying students lends support to the DIBELS team's decision to stop recommending PSF for use as a screener in first grade. As with NWF, though the cut scores obtained in this study were considerably higher than those set in the DIBELS manual, they were similar to those found by other researchers. Despite these higher cut scores, however, specificity still suffered. For fall PSF, Johnson et al (2009) used a cut score of 53, which resulted in specificity of .22, a value even lower than that found in this study. Clemens et al. (2011) found almost the same results. They used a cut score of 50 and obtained specificity equal to .20.

Language proficiency. One of the unique contributions of this study was to run ROC analyses separately by English language proficiency group. Interestingly, the cut scores needed to obtain a sensitivity near .90 were almost identical for the B/EI group, I/EA group, and the total sample. The only exception was the cut score for fall NWF for the I/EA group. For this group, the derived cut score was about half that of the total sample and B/EI group. If the higher cut score had been used with this group, 16 more students (out of 51) would have been falsely identified as at-risk. While that is not as bad as overlooking students who were truly at-risk, it could cause a serious strain on a school's resources to intervene unnecessarily with 30% of their students.

From this evidence, would it be appropriate to recommend that teachers use a lower cut score for their advanced ELLs? At this point, such a decision would be premature. There are no comparable findings to show that this is a consistent result. On the contrary, the NWF studies with NESs varied considerably in the cut scores identified (i.e., 23-38 in the fall, and 24-50 in the winter), so this may be a characteristic of the measure itself and its sensitivity to growth rates.

Best Screener

As multi-tiered support systems have become increasingly common in schools, experts have highlighted key characteristics of high-quality literacy screeners (Fuchs & Fuchs, 1998; Hayes, Nelson & Jarrett, 1987; Reschly & Wilson, 1995). They need to be technically adequate, of course, and should also be able to predict future performance and accurately classify students as at-risk or not at-risk. In addition, screeners should provide information that will guide school staff in determining how to intervene with struggling students, and screeners need to be easy to administer to large numbers of students. This study focused on four screeners, their ability to predict end-of-year performance on ORF, and how accurately they could classify students based on risk level.

After reviewing the results from each of the different analyses, the WIF screener was clearly the strongest predictor of end-of-first grade ORF scores. It was the most strongly correlated with ORF3 in fall, winter, and spring. WIF was the only significant predictor of ORF3 in the winter regression model, and one of two (with NWF) significant predictors in the fall. When it comes to classification, WIF had the largest AUC of the fall and winter measures and the highest specificity (when sensitivity was held at .90). So, the findings from this study support other researchers' recommendations to use WIF as a screener for first grade students (e.g., Clemens et al., 2011; Han et al., 2015).

The DST measure was included in this study to explore its utility as a possible screener. It was hypothesized that it would predict better than PSF, which it did. In fact, it exceeded expectations by predicting almost as well as NWF and classifying better. DST was strongly correlated with ORF3 at each period and was significant in each regression model until WIF was added. In the fall, the AUC for DST was equal to NWF's, and in the winter, it was larger than NWF's. DST also had higher specificity than NWF (when sensitivity was held at .90) at both time periods. While research into spelling as a screener is still emerging, this study lends further support to earlier recommendations for schools to seriously consider using it (Chua et al., 2016; Clemens et al., 2014; Munger & Murray, 2017).

However, these findings should not be considered sufficient evidence to throw out PSF or NWF. For example, though PSF is a weak predictor, prediction is not the only role of screeners. Within a multi-tiered support system, they frequently play the role of a formative assessment. Though this study only examined prediction and classification, teachers need assessment tools that will help them identify and address students' specific, immediate needs (Gersten et al., 2008), as much or even more than they need assessments to help them predict long-term performance. Regardless of PSF's ability to predict later reading scores, nonreaders still need instruction in phonological awareness (NICHHD, 2000), and educators would do well to keep that in mind. Both PSF and NWF have been used because they are closely connected with important intervention targets.

Examined through this lens, WIF has a weakness. Despite its strength as a predictor, WIF does not help teachers decide where to intervene. Because it measures multiple early literacy skills, it can be difficult to diagnose the reason for a low score.

Does the student have a problem with phonological awareness, decoding, or fluency? Which skill should be taught?

DST holds more promise in this area. There is evidence from other studies that teaching invented spelling can be a useful intervention and one that increases students' subsequent PA and phonics skills (Levin & Aram, 2013; O'Connor & Jenkins, 1995; Sénéchal, Ouellette, Pagan, & Lever, 2012). DST also stands out from the other screeners when considering ease of administration. With DST, one could test the entire class at once rather than giving the test one-on-one. Unfortunately, this ease of administration may be counterbalanced initially by the difficulty of scoring the assessment (Munger & Murray, 2017), but that problem would be mitigated over time as teachers become more experienced and proficient at grading the test.

Impact of Language Proficiency

In addition to examining which screeners best predict fluency at the end of first grade, this study was designed to examine whether English language proficiency affects the screeners' ability to predict future performance. PPVT was included as a continuous measure that served as a proxy for language proficiency, and PPVT was significantly correlated with all the screeners and the outcome measure. However, when correlations were disaggregated by CELDT level, there were no significant differences between results for the total sample, B/EI group, or I/EA group. Similarly, neither the CELDT nor PPVT variables explained additional variance in ORF3 when added to the hierarchical regression model after the literacy screeners. When ROC analyses were run, the screeners that were the strongest for the total sample were strongest for the ELL groups as well, and when specific cut scores were examined, only winter NWF for the I/EA group had a noticeably different cut score.

Practical Implications

What can be concluded from these findings? Within a multi-tiered support system (MTSS), the goal of identifying the best screener with the right cut score is to accurately identify and intervene with at-risk students so students can be successful. The findings in this study indicate that one does not need to consider language proficiency level when examining screening scores. This is good news for teachers as it simplifies the process and allows them to use and interpret the same screeners for the Ss ELLs in their class as for their other students.

Though WIF and DST are not part of the commonly used DIBELS battery, the results from this study indicate that schools should consider using both screeners in first grade. WIF does an excellent job predicting future reading performance, and DST may serve as a useful replacement for PSF at the first-grade level. However, one word of warning is appropriate here. In the business world, there is a saying: “What gets measured gets done.” Unfortunately, in education that may mean teaching to the test. In their seminal 1982 study of curriculum-based measurement, Deno and colleagues found that using isolated word lists, like WIF, accidentally encouraged teachers to teach isolated words. In their words, “Nothing about the research reported here supports practice on reading words aloud as an effective approach to improving reading proficiency” (p. 44). The same caution is applicable here. As with all screeners, any school that chooses to use WIF should make the purpose explicit, and specifically in the case of this assessment, school staff should be aware that it is a diagnostic tool, not a list of words students need to memorize.

Limitations

However, a few limitations should be considered in drawing conclusions from this data. The decision to remove students with incomplete data from the sample may have influenced the results. Also, the cut scores evaluated were sample-dependent and have not been cross-validated with other samples, so they may not be generalizable.

In addition, the measures used to assess ELLs' language proficiency were not ideal. PPVT is a measure of receptive vocabulary, not language proficiency in general. Though vocabulary is often the area where ELLs are most different from NESs, this measure was not designed for use with ELLs and may not have been broad enough to pick up all the differences between the two groups. On the other hand, though the CELDT is a broader measure, the available data consisted of categorical scores, not continuous. Also, none of the students in the current sample scored at the highest level, so this restricted range may have made it more difficult to identify the impact of their degree of language proficiency.

Other differences among ELLs, such as country of origin or SES, were not considered since those data were not available. Prior research indicates that those factors may play a role and be worth considering (Halle et al., 2012; Palardy, 2015).

One final limitation is applicable to all screening studies. This type of study has no way to evaluate whether classification errors were due to problems with the screeners themselves or to the effectiveness of the instruction students received. All students will be receiving instruction, of course, but without any data evaluating the quality of instruction, researchers cannot identify studies where the screener appeared poor because of variation in instructional quality. Highly effective instruction would tend to cause an increase in false positives and, consequently, a decrease in specificity. In this study, for example, 30 students received intervention as part of the larger research study. Of those students, 18 scored below the 25th percentile on ORF at the end of the year. Using their scores on the fall screeners, these would all be true positives which yields a sensitivity of 100%. However, 12 students improved sufficiently that they scored above the cut score. This results in specificities of 8%, 42%, 8%, and 0% on PSF, DST, NWF, and WIF, respectively. Though these specificity levels would typically be used as evidence of poor classification accuracy, in this case they are clearly an indication of successful intervention.

Future Research

There are several different avenues for future research that would build on this study's findings. For example, though this study focused on the predictive power and classification accuracy of the four screeners included, screeners play additional roles. Future research should examine how teachers and administrators use screening results. Are they more concerned with prediction or with diagnosis? Similarly, future research should consider how well DST and WIF assist in the problem-solving process that is part of a MTSS. Do teachers make better decisions about intervention with these screeners than with PSF and NWF?

When it comes to screening studies, spelling is still a fairly new skill to evaluate in first grade. Researchers should consider piloting the use of spelling as a screener and getting teacher feedback on its usefulness and whether it is easy to administer and score. Researchers could also examine whether DST is an appropriate measure for progress monitoring. At the first-grade level, researchers could compare different types of spelling assessments as Clemens and colleagues (2014) did at the kindergarten level. Such research could help move toward consensus on one particular type of spelling screener, which would then allow for replication and comparability across studies.

At this point, the research seems consistent for Ss ELLs - the same early literacy screeners can be used for them as for NESs. However, it is important to note research has reported differences for ELLs from other language backgrounds (e.g., Han et al., 2015). Researchers who are studying students from other language backgrounds should keep this in mind and continue to consider this issue. Also, future research should examine whether relations differ for Ss ELLs in later grades. Research indicates, for example, that language proficiency may play a larger role later in elementary school when vocabulary becomes more important than word reading skills for reading comprehension (Gottardo & Mueller, 2009).

References

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

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

August, D. & Shanahan, T. (Eds.). (2006). *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Beach, K. D., & O'Connor, R. E. (2015). Early response-to-intervention measures and criteria as predictors of reading disability in the beginning of third grade. *Journal of Leanring Disabilities, 48*(2), 196-223. doi:10.1177/0022219413495451

Burke, M. D., & Hagen-Burke, S. (2007). Concurrent criterion-related validity of early literacy indicators for the middle of first grade. *Assessment for Effective Intervention, 32*(2), 66-77.

Burke, M. D., Hagen-Burke, S., Kwok, O., & Parker, R. (2009). Predictive validity of early literacy indicators from the middle of kindergarten to second grade. *Journal of Special Education, 42*(4), 209-226.

California Department of Education. (2003). *Decision Guide: Reclassifying a Student from EL to FEP.* Retrieved from http://celdt.cde.ca.gov/CELDT2003/Celdt/reclass_EL_03.pdf

California Department of Education. (2013). *CELDT Number and Percent of Students at Each Overall Performance Level.* Retrieved from http://dq.cde.ca.gov/dataquest/CELDT/results.aspx?year=2011-2012&level=state&assessment=3&subgroup=1&entity=

California Department of Education. (2016). *CalEdFacts.* Retrieved from http://www.cde.ca.gov/re/pn/fb/index.asp

California Department of Education. (2017). DataQuest Student & School Data Reports. Retrieved from http://dq.cde.ca.gov/dataquest/dataquest.asp

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.

Chua, S. M., Liow, S. J. R. & Yeong, S. H. M. (2016). Using spelling to screen bilingual kindergarteners at risk for reading difficulties. *Journal of Learning Disabilities, 49*(3), 227-239.

Clemens, N. H. (2009). *Toward consensus on first grade CBM measures* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3373071)

Clemens, N. H., Hilt-Panahon, A., Shapiro, E. S., & Yoon, M. (2012). Tracing student responsiveness to intervention with early literacy skills indicators: Do they reflect growth toward text reading outcomes? *Reading Psychology, 33*, 47-77

Clemens, N. H., Oslund, E. L., Simmons, L. E., & Simmons, D. (2014). Assessing spelling in kindergarten: Further comparison of scoring metrics and their relation to reading skills. *Journal of School Psychology*, *52*(1), 49-61. doi:10.1016/j.jsp.2013.12.005

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.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edition.* Hillsdale, N.J.: Lawrence Erlbaum.

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

Compton, D. L. (2000). Modeling the growth of decoding skills in first-grade children. *Scientific Studies of Reading, 4,* 219-259.

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. doi:10.1037/0022-0663.98.2.394

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., … Crouch, R. C. (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. doi:10.1037/a0018448

CTB/McGraw-Hill LLC. (2008). *Technical Report for the California English Language Development Test (CELDT)*. Retrieved from http://www.cde.ca.gov/ta/tg/el/documents/techrpt0708.pdf

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

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

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

Dunn, L. M., & Dunn, D. M. (1997). The Peabody Picture Vocabulary Test. Circle Pines, MN: American Guidance Services.

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.

Fletcher, J. M, Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2007). *Learning disabilities: From identification to intervention.* New York, NY: Guilford Press.

Ford, K. L., Cabell, S. Q., Konold, T. R., Invernizzi, M., & Gartland, L. B. (2013). Diversity among Spanish-speaking English language learners: Profiles of early literacy skills in kindergarten. *Reading & Writing, 26*, 889-912. doi:10.1007/s11145-012-9397-0

Fuchs, L. S. & Fuchs, D. (2007). *Using curriculum-based measurement for progress monitoring in reading*. Retrieved from http://www.eric.ed.gov/PDFS/ED519252.pdf

Fuchs, L. S., & Fuchs, D. (2004). Determining adequate yearly progress from kindergarten through grade 6 with curriculum-based measurement. *Assessment for Effective Intervention, 29*(4), 25-37.

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

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.

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

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

Goffreda, C. T., DiPerna, J. C., & Pedersen, J. A. (2009). Preventive screening for early readers: Predictive validity of the Dynamic Indicators of Basic Early Literacy Skills. *Psychology in the Schools, 46*(6), 539-552. doi:10.1002/pits.20396

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

Good, R. H., Kaminski, R. A., Shinn, M., Bratten, J., Shinn, M., Laimon, D., . . . Flindt, N. (2004). *Technical adequacy of DIBELS: Results of the Early Childhood Research Institute on measuring growth and development* (Technical Report, No. 7). Eugene, OR: University of Oregon.

Goodrich, J. M., Farrington, A. L., & Lonigan, C. J. (2016). Relations between early reading writing skills among Spanish-speaking language minority children. *Reading & Writing, 29*, 297-319. doi:10.1007/s11145-015-9594-8

Gottardo, A., & Mueller, J. (2009). Are first- and second-language factors related in predicting second-language reading comprehension? A study of Spanish-speaking children acquiring English as a second language from first to second grade. *Journal of Educational Psychology, 101*(2), 330-344. doi:10.1037/a0014320

Gutiérrez, G. (2010). *An examination of the quality of literacy skill assessments across levels of second-grade, Spanish-speaking, English-language learners* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3412781)

Halle, T., Hair, E., Wandner, L., McNamara, M., & Chien, N. (2012). Predictors and outcomes of early versus later English language proficiency among English language learners. *Early Childhood Research Quarterly, 27,* 1-20.

Han, J. N., Vanderwood, M. L., & Lee, C. Y. (2015). Predictive validity of early literacy measures for Korean English language learners in the United States. *International Journal of School & Educational Psychology, 3*(3), 178-188. doi:10.1080/21683603.2015.1059781

Harrison, G. L., Goegan, L. D., Jalbert, R., McManus, K., Sinclair, K., & Spurling, J. (2016). Predictors of spelling and writing skills in first- and second-language learners. *Reading & Writing, 29,* 69-89. doi:10.1007/s11145-015-9580-1

Hayes, S.C., Nelson, R.O., & Jarrett, R.B. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist, 42,* 963-974.

Healy, K. D. (2007). *Word identification fluency and nonsense word fluency as predictors of reading fluency in first grade.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3298219)

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.

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., & 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*(4), 437-447.

Kame’enui, E. J. (2002). *Executive summary of final report on Reading First reading assessment analysis*. Retrieved from AIMSweb: http://www.aimsweb.com

Kame'enui, E. J. & Simmons, D. C. (2001). Introduction to this special issue: The DNA of reading fluency. *Scientific Studies of Reading, 5*(3), 203-210.

Kaminski, R. A., & Good, R. H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review, 25*(2), 215.

Kim, J. S. (2012). *Examining the predictive validity of DIBELS literacy measures with third grade Spanish-speaking English language learners.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3503282)

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.

Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology, 95*(1), 3-21. doi:10.1037/0022-0663.95.1.3

Kupzyk, S., Daly, E. J., Ihlo, T., & Young, N. D. (2012). Modifying instruction within tiers in multitiered intervention programs. *Psychology in the Schools*, *49*(3), 219-230. doi:10.1002/pits.21595

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

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, NY: Routledge.

Lesaux, N. & Geva, E. (2006). Synthesis: Development of literacy in language-minority students. In D. August, & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on Language Minority Children and Youth* (pp. 53-74). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Levin, I. & Aram, D. (2013). Promoting early literacy via practicing invented spelling: A comparison of different mediation routines. *Reading Research Quarterly, 48*(3), 221-235. doi:10.1002/rrq.48

Linan-Thompson, S., & Vaughn, S. (2010). Evidence-based reading instruction: Developing and implementing reading programs at the core, supplemental, and intervention levels. In G. G. Peacock, R. A. Ervin, E. J. Daly, & K. W. Merrell (Eds.), *Practical handbook of school psychology: Effective practices for the 21st century* (pp. 274-286). New York, NY: Guilford.

Linan-Thompson, S., Cirino, P. T., & Vaughn, S. (2007). Determining English language learners' response to intervention: Questions and some answers. *Learning Disability Quarterly, 30*, 185-195.

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

Lindsey, K. A., Manis, F. R., & Bailey, C. E. (2003). Prediction of first-grade reading in Spanish-speaking English-language learners. *Journal of Educational Psychology, 95*(3), 482-494. doi:10.1037/0022-0663.95.3.482

Linklater, D. L., O'Connor, R. E., & Palardy, G. J. (2009). Kindergarten literacy assessment of English Only and English language learner students: An examination of the predictive validity of three phonemic awareness measures. *Journal of School Psychology, 47*, 369-394.

Mancilla-Martinez, J., & Lesaux, N. K. (2010). Predictors of reading comprehension for struggling readers: The case of Spanish-speaking language minority learners. *Journal of Educational Psychology, 102*(3), 701-711. doi:10.1037/a0019135

Manis, F. R., Lindsey, K. A., & Bailey, C. E. (2004). Development of reading in grades K-2 in Spanish-speaking English-language learners. *Learning Disabilities Research & Practice, 19*(4), 214-224.

Mesmer, H. A. E., & Williams, T. O. (2014). Modeling first grade reading development. *Reading Psychology, 35*(5), 468-495. doi:10.1080/02702711.2012.743494

Morris, D., Trathen, W., Perney, J., Gill, T., Schlagal, R., Ward, D. & Frye, E. M. (2017). Three DIBELS tasks vs. three informal reading/spelling tasks: A comparison of predictive validity. *Reading Psychology, 38*(3), 289-320. doi:10.1080/02702711.2016.1263700

Munger, K. A., & Blachman, B. A. (2013). Taking a 'simple view' of the Dynamic Indicators of Basic Early Literacy Skills as a predictor of multiple measures of third-grade reading comprehension. *Psychology in the Schools, 50*(7), 722-737. doi:10.1002/pits.21699

Munger, K. A., & Murray, M. S. (2017). First-grade spelling scores within the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) screening: An exploratory study. *Educational Assessment, 22*(2), 124-137. doi:10.1080/10627197.2017.1309275

Nam, J. E. (2011). *An examination of the predictive validity of early literacy measures for Korean English language learners.* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3491308)

National Association of State Directors of Special Education. (2007). *Response to Intervention research for practice*. Retrieved from http://http://www.nasdse.org/

National Center for Education Statistics. (2011). *The Nation's Report Card: Reading 2011* (NCES 2012–457). Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

National Center on Response to Intervention. (2010). *Essential components of RTI: A closer look at Response to Intervention.* Retrieved from http://www.rti4success.org/pdf/rtiessentialcomponents_04271.pdf

National Clearinghouse for English Language Acquisition. (2010). *The growing number of English learner students.* NCELA Publications. Washington, DC: Author. Retrieved from http://www.ncela.gwu.edu/files/uploads/9/growing_EL_0910.pdf

National Clearinghouse for English Language Acquisition. (2011). *What languages do English learners speak?* NCELA Fact Sheet. Washington, DC: Author. Available from www.gwu.edu/files/uploads/NCELAFactsheets/EL_Languages_2011.pdf

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.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

O'Connor, R. E., Bocian, K., Beebe-Frankenberger, M., & Linklater, D. L. (2010). Responsiveness of students with language difficulties to early intervention in reading. *Journal of Special Education*, *43*(4), 220-235. doi:10.1177/0022466908317789

O'Connor, R. E., Bocian, K. M., Sanchez, V., & Beach, K. D. (2014). Access to a responsiveness to intervention model: does beginning intervention in kindergarten matter? *Journal of Learning Disabilities, 47*(4), 307-328. doi:10.1177/0022219412459354

O'Connor, R. E., & Jenkins, J. R. (1995). Improving the generalization of sound/symbol knowledge: Teaching spelling to kindergarten children with disabilities. *Journal of Special Education, 29*(4), 255-275.

O'Connor, R. E., & Jenkins, J. R. (1999). Prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading, 3*(2), 159-197. doi:10.1207/s1532799xssr0302_4

Palardy, G. (2015). Classroom-based inequalities and achievement gaps in first grade: The role of classroom context and access to qualified and effective teachers. *Teachers College Record, 117*, 1-48.

Pearson Clinical. (2017). Peabody Picture Vocabulary Test-Third Edition (PPVT-III). Retrieved from http://www.pearsonclinical.com/language/products/100000081/peabody-picture-vocabulary-test-third-edition-ppvt-iii.html#tab-scoring

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (Third ed.). Fort Worth, TX: Wadsworth/Thomson Learning.

Quiroga, T., Lemos-Britton, Z., Mostafapour, E., Abbott, R. D., & Berninger, V. W. (2002). Phonological awareness and beginning reading in Spanish-speaking ESL first graders: Research into practice. *Journal of School Psychology, 40*(1), 85–111.

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

Reschly, D.J., & Wilson, M. S. (1995). School psychology practitioners and faculty: 1986 to 1991–92—Trends in demographics, roles, satisfaction, and system reform. *School Psychology Review*, 24, 62–80.

Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42*(4), 546-567. doi:10.1598/RRQ.42.4.5

Ryan, C. L., & Bauman, K. (2016). *Educational attainment in the United States: 2015* (Report No. P20-578). Retrieved from United States Census Bureau website: http://www.census.gov/content/dam/Census/library/publications/2016/demo/p20-578.pdf

Salvia, J., Ysseldyke, J. E., & Bolt, S. (2010). *Assessment in special and inclusive education* (11th ed.). Belmont, CA: Wadsworth.

Scanlon, D. M., Vellutino, F. R., Small, S. G., Fanuele, D. P., & Sweeney, J. M. (2005). Severe reading difficulties—can they be prevented? A comparison of prevention and intervention approaches. *Exceptionality*, 13(4), 209-227. doi:1.1207/s15327035ex1304_3

Sénéchal, M., Ouellette, G., Pagan, S., & Lever, R. (2012). The role of invented spelling on learning to read in low-phoneme-awareness kindergartners: A randomized-control-trial study. *Reading and Writing: An Interdisciplinary Journal*, *25*, 917–934. doi:10.1007/s11145-011-9310-2

Shin, H.B. and Kominski, R.A. (2010). *Language Use in the US: 2007.* ACS Reports, ACS-12. Washington, DC: US Census Bureau. Retrieved from http://www.census.gov/hhes/socdemo/language/data/acs/ACS-12.pdf

Slocum, T. A., O'Connor, R. E., & Jenkins, J. R. (1993). Transfer among phonological manipulation skills. *Journal of Educational Psychology, 85*(4), 618-630.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academies Press.

Solari, E. J., Aceves, T. C., Higareda, I., Richards-Tutor, C., Filippini, A. L., Gerber, M. M., Leafstedt, J. (2014). Longitudinal prediction of 1st and 2nd grade English oral reading fluency in English language learners: Which early reading and language skills are better predictors? *Psychology in the Schools, 51*(2), 126-142. doi:10.1002/pits.21743

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

Stanovich, K. E., & Siegel, L. S. (1994). Phenotypic performance profile of children with reading disabilities: A regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology, 86,* 24–53.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285-1293. doi:10.1126/science.3287615

Tangel, D. M., & Blachman, B. A. (1992). Effect of phoneme awareness instruction on kindergarten children's invented spelling. *Journal of Reading Behavior, 24*(2), 233–261.

Tangel, D. M., & Blachman, B. A. (1995). Effect of phoneme awareness instruction on the invented spelling of first-grade children: A one-year follow-up. *Journal of Reading Behavior, 27,* 153–185. doi:10.1080/10862969509547876

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Burgess, S., & Hecht, S. (1997). Contributions of phonological awareness and rapid automatic naming ability to the growth of word-reading skills in second to fifth-grade children. *Scientific Studies of Reading, 1,* 161–185.

University of Oregon Center on Teaching & Learning. (2012). *2012–2013 DIBELS Data System Update Part II: DIBELS Next Benchmark Goals* (Technical Brief No. 1203). Eugene, OR: University of Oregon. Retrieved from https://dibels.uoregon.edu/docs/techreports/DDS2012TechnicalBriefPart2.pdf

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.

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. E., & Sun, J. W. (2014). Validity of DIBELS early literacy measures with Korean English learners. *Contemporary School Psychology, 18,* 205-213. doi:10.1007/s40688-014-0032-8

Vellutino, F. R., Tunmer, W. E., Jaccard, J. J. & Chen, R. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading, 11*(1), 3-32. doi:10.1080/10888430709336632

White House Initiative on Educational Excellence for Hispanics, & U.S. Department of Education. (2011). *Winning the future: Improving education for the Latino community.* Retrieved from http://www.whitehouse.gov/sites/default/files/rss_viewer/WinningTheFutureImprovingLatinoEducation.pdf

Zumeta, R. O., Compton, D. L., & Fuchs, L. S. (2012). Using Word Identification Fluency to monitor first-grade reading development. *Exceptional Children, 78*(2), 201-220.

Table 1

Student Demographics

			Ethnicity					
	Enrollment	ELLs	Hispanic	AA	White	All Others	Multiple / DNR	FRL
District A	56,727	33%	68%	16%	11%	2%	3%	81%
School 1	910	34%	63%	18%	12%	2%	4%	85%
School 2	406	34%	69%	16%	9%	0%	5%	95%
School 3	475	33%	65%	18%	10%	3%	4%	88%
District B	19,987	45%	72%	4%	15%	6%	3%	53%
School 4	477	60%	80%	4%	12%	1%	4%	67%
School 5	601	52%	76%	4%	13%	6%	1%	52%

Note. ELLs = English language learners; AA = African American; DNR = Did not respond; FRL = Free and reduced lunch.

Table 2

Predictive Accuracy Formulas

	At-Risk on ORF	Not At-Risk on ORF	
At-Risk on Screener	True positives (TP)	False positives (FP)	Positive predictive power = TP/(TP + FP)
Not At-Risk on Screener	False negatives (FN)	True negatives (TN)	Negative predictive power = TN/(TN + FN)
	Sensitivity = TP/(TP + FN)	Specificity = TN/(TN + FP)	Hit rate = (TP + TN)/(TP + TN + FP + FN)

Table 3

Descriptive Statistics for Entire Sample (N = 209), B/EI (n = 55), and I/EA (n = 51) Groups

Measure	Range			*M (SD)*		
Fall	Overall	B/EI	I/EA	Overall	B/EI	I/EA
PSF1	0-85	0-69	7-70	38.24 (18.42)	31.32 (20.21)	41.84 (16.90)
DST1	5-48	12-48	22-48	39.61 (7.14)	37.62 (8.04)	41.39 (5.23)
NWF1	0-116	0-116	4-71	29.95 (16.67)	25.71 (17.38)	32.18 (16.05)
WIF1	0-69	0-49	0-61	12.98 (12.28)	9.07 (9.53)	16.18 (13.33)
PPVT	45-120	45-102	50-111	85.71 (12.13)	77.18 (11.62)	85.16 (10.39)
Winter						
PSF2	6-109	12-70	17-88	49.04 (13.78)	48.17 (12.17)	47.43 (13.08)
DST2	17-59	17-59	29-59	50.61 (6.51)	49.62 (7.24)	50.69 (6.45)
NWF2	7-129	7-89	10-129	46.09 (19.98)	40.73 (16.61)	51.02 (26.51)
WIF2	0-105	0-82	6-105	33.34 (20.30)	28.82 (18.57)	38.63 (21.16)
Spring						
DST3	18-72	18-70	51-71	60.96 (6.77)	59.71 (8.53)	62.37 (4.88)
NWF3	0-145	0-145	5-127	52.90 (24.97)	48.53 (25.14)	55.20 (24.40)
WIF3	0-100	0-97	10-93	42.40 (21.21)	39.20 (20.15)	47.82 (20.12)
ORF3	0-122	0-107	7-114	44.96 (23.61)	39.85 (22.31)	51.75 (22.93)

Note. B/EI = Beginning/Early Intermediate; I/EA = Intermediate/Early Advanced; PSF = Phoneme Segmentation Fluency; DST = Developmental Spelling Test; NWF = Nonsense Word Fluency; WIF = Word Identification Fluency; PPVT = Peabody Picture Vocabulary Test; ORF = Oral Reading Fluency.

Table 4

Correlations Between Measures for Entire Sample (N = 209)

	PSF1	DST1	NWF1	WIF1	PSF2	DST2	NWF2	WIF2	DST3	NWF3	WIF3	PPVT	ORF3
PSF1													
DST1	.39**												
NWF1	.45**	.48**											
WIF1	.35**	.46**	.66**										
PSF2	.38**	.26**	.28**	.09									
DST2	.30**	.73**	.46**	.38**	.26**								
NWF2	.26**	.43**	.52**	.50**	.25**	.47**							
WIF2	.30**	.57**	.60**	.71**	.14*	.60**	.66**						
DST3	.33**	.69**	.39**	.36**	.19**	.79**	.39**	.57**					
NWF3	.22**	.39**	.50**	.47**	.17*	.44**	.65**	.61**	.42**				
WIF3	.23**	.55**	.55**	.65**	.17*	.61**	.60**	.88**	.62**	.65**			
PPVT	.23**	.32**	.19**	.22**	.23**	.32**	.17*	.21**	.22**	.15*	.19**		
ORF3	.28**	.56**	.58**	.72**	.16*	.60**	.67**	.91**	.61**	.67**	.91**	.20**	

Note. PSF = Phoneme Segmentation Fluency; DST = Developmental Spelling Test; NWF = Nonsense Word Fluency; WIF = Word Identification Fluency; PPVT = Peabody Picture Vocabulary Test; ORF = Oral Reading Fluency
**p* < .05 ***p* < .01

Table 5

Correlations Between Measures for B/EI (bottom; n = 55) and I/EA (top; n = 51) Groups

	PSF1	DST1	NWF1	WIF1	PSF2	DST2	NWF2	WIF2	DST3	NWF3	WIF3	PPVT	ORF3
PSF1	-	.18	.43**	.39**	.40**	.07	.15	.22	.15	.14	.27	.11	.24
DST1	.44**	-	.46**	.48**	.29*	.74**	.39**	.51**	.74**	.34*	.53**	.44**	.57**
NWF1	.41**	.49**	-	.59**	.48**	.48**	.52**	.56**	.46**	.56**	.58**	.13	.64**
WIF1	.25	.49**	.73**	-	.19	.39**	.54**	.63**	.47**	.50**	.69**	.27	.70**
PSF2	.27*	.16	.12	-.01	-	.30*	.28	.19	.30*	.27	.25	.32*	.19
DST2	.33*	.80**	.47**	.42**	.39**	-	.50**	.57**	.76**	.47**	.57**	.34*	.59**
NWF2	.15	.49**	.53**	.61**	.10	.47**	-	.66**	.36**	.79**	.69**	.33*	.73**
WIF2	.22	.60**	.65**	.77**	.08	.60**	.69**	-	.58**	.60**	.90**	.31*	.90**
DST3	.33*	.70**	.41**	.40**	.25	.86**	.41**	.58**	-	.37**	.58**	.37**	.62**
NWF3	.20	.48**	.56**	.74**	.05	.46**	.78**	.75**	.48**	-	.66**	.23	.59**
WIF3	.19	.58**	.63**	.73**	.23	.67**	.65**	.90**	.67**	.73**	-	.32*	.91**
PPVT	.33*	.30*	.08	.08	.30*	.37**	.03	.08	.32*	.01	.09	-	.35*
ORF3	.16	.57**	.57**	.76**	.10	.60**	.72**	.93**	.60**	.78**	.93**	.04	-

Note. B/EI = Beginning/Early Intermediate; I/EA = Intermediate/Early Advanced; PSF = Phoneme Segmentation Fluency; DST = Developmental Spelling Test; NWF = Nonsense Word Fluency; WIF = Word Identification Fluency; PPVT = Peabody Picture Vocabulary Test; ORF = Oral Reading Fluency.
**p* < .05 ***p* < .01

Table 6

Hierarchical Regression Analyses with Fall Screeners Predicting Spring ORF

		R^2	Adj R^2	ΔR^2	Final *B*	Final β	*F*
Step 1:	PSF1	.08	.08	.08**	-0.08	-.07	17.99**
Step 2:	DST1	.38	.37	.30**	-0.35	-.11	41.15**
	DST1^2				0.02	.42	
Step 3:	NWF1	.48	.47	.10**	0.18	.12	46.62**
Step 4:	WIF1	.60	.59	.12**	0.96	.50**	61.03**
Step 5a:	LangProf	.61	.59	.00			43.68**
	B/EI				2.67	.05	
	I/EA				3.27	.06	
Step 5b:	PPVT	.60	.59	.00	-0.04	-.02	50.68**

Note. ORF = Oral Reading Fluency; PSF = Phoneme Segmentation Fluency; DST = Developmental Spelling Test; NWF = Nonsense Word Fluency; WIF = Word Identification Fluency; LangProf = Language Proficiency; B/EI = Beginning/Early Intermediate; I/EA = Intermediate/Early Advanced; PPVT = Peabody Picture Vocabulary Test.
*p < .05 **p < .001

Table 7

Hierarchical Regression Analyses with Winter Screeners Predicting Spring ORF

		R^2	Adj R^2	ΔR^2	Final *B*	Final β	*F*
Step 1:	PSF2	.03	.02	.03*	0.01	.01	5.73*
Step 2:	DST2	.45	.45	.43**	-0.51	-.14	56.67**
	DST2^2				0.01	.23	
Step 3:	NWF2	.60	.60	.15**	0.13	.11**	77.27**
Step 4:	WIF2	.84	.84	.24**	0.90	.78**	220.05**
Step 5a:	LangProf	.85	.84	.00			154.18**
	B/EI				0.71	.01	
	I/EA				2.14	.04	
Step 5b:	PPVT	.84	.84	.00	-0.01	-.01	182.51**

Note. ORF = Oral Reading Fluency; PSF = Phoneme Segmentation Fluency; DST = Developmental Spelling Test; NWF = Nonsense Word Fluency; WIF = Word Identification Fluency; LangProf = Language Proficiency; B/EI = Beginning/Early Intermediate; I/EA = Intermediate/Early Advanced; PPVT = Peabody Picture Vocabulary Test.
*p < .05 **p < .01

Table 8

Hierarchical Regression Analyses with Fall Screeners Predicting Winter WIF

		R^2	Adj R^2	ΔR^2	Final *B*	Final β	*F*
Step 1:	PSF1	.09	.08	.09**	-0.07	-.06	19.79**
Step 2:	DST1	.39	.38	.30**	-0.61	-.21	43.54**
	DST1^2				0.02	.53*	
Step 3:	NWF1	.50	.49	.12**	0.21	.17**	51.76**
Step 4	WIF1	.61	.60	.10**	0.75	.45**	62.40**

Note. WIF = Word Identification Fluency; PSF = Phoneme Segmentation Fluency; DST = Developmental Spelling Test; NWF = Nonsense Word Fluency.
**p* < .05 ***p* < .01

Table 9

Area Under the Curve and Maximized Sensitivity and Specificity for PSF, DST, NWF, and WIF on Below the 25th Percentile (ORF3 < 28) of Spring ORF (N = 209)

			Max		
Measure	AUC	95% CI	*Sn*	*Sp*	Cut Score
Fall					
PSF1	.69	[.61, .77]	.61	.75	34
DST1	.84	[.78, .91]	.75	.81	40
NWF1	.84	[.77, .90]	.71	.85	21
WIF1	.87	[.82, .92]	.82	.76	8
Winter					
PSF2	.63	[.54, .71]	.45	.73	43
DST2	.87	[.81, .92]	.90	.67	52
NWF2	.81	[.75, .88]	.88	.62	44
WIF2	.97	[.96, .99]	.92	.92	19

Note. PSF = Phoneme Segmentation Fluency; DST = Developmental Spelling Test; NWF = Nonsense Word Fluency; WIF = Word Identification Fluency; ORF = Oral Reading Fluency; AUC = Area under the curve; Sn = Sensitivity; Sp = Specificity.

Table 10

Predictive Accuracy Indices for PSF, DST, NWF, and WIF on Below the 25th Percentile (ORF3 < 28) of Spring ORF (N = 209)

Measure	AUC	95% CI	Cut Score*	*Sn*	*Sp*	TP	FP	TN	FN	PPP	NPP	Hit rate
Fall												
PSF1	.69	[.61, .77]	49	.90	.36	46	101	57	5	.32	.92	.49
DST1	.84	[.78, .91]	42	.88	.60	45	64	94	6	.41	.94	.67
NWF1	.84	[.77, .90]	34	.90	.46	46	86	72	5	.35	.94	.56
WIF1	.87	[.82, .92]	10	.90	.65	46	55	103	5	.46	.95	.71
Winter												
PSF2	.63	[.54, .71]	60	.90	.25	46	119	39	5	.28	.89	.41
DST2	.87	[.81, .92]	52	.90	.67	46	52	106	5	.47	.95	.73
NWF2	.81	[.75, .88]	49	.90	.48	46	82	76	5	.36	.94	.48
WIF2	.97	[.96, .99]	19	.92	.92	47	12	146	4	.80	.97	.92

Note. PSF = Phoneme Segmentation Fluency; DST = Developmental Spelling Test; NWF = Nonsense Word Fluency; WIF = Word Identification Fluency; ORF = Oral Reading Fluency; AUC = Area under the curve; Sn = Sensitivity; Sp = Specificity; TP = True positive; FP = False positive; TN = True negative; FN = False negative; PPP = Positive predictive power; NPP = Negative predictive power.
*Cut scores were identified that resulted in sensitivity levels as close to .90 as possible.

Table 11

Predictive Accuracy Indices by ELL Group for Fall Screeners on Below 25th Percentile of Spring ORF

Measure	AUC*	95% CI	Cut Score*	Sn	Sp	TP	FP	TN	FN	PPP	NPP	Hit rate
B/EI (*n* = 55)												
PSF1	.65	[.48, .81]	49	.93	.28	14	29	11	1	.33	.92	.45
DST1	.78	[.63, .93]	42	.93	.48	14	21	19	1	.40	.95	.60
NWF1	.81	[.68, .94]	30	.93	.38	14	25	15	1	.36	.94	.53
WIF1	.85	[.74, .95]	9	.93	.50	14	20	20	1	.41	.95	.62
I/EA (*n* = 51)												
PSF1	.60	[.40, .80]	50	.88	.37	7	27	16	1	.21	.94	.45
DST1	.77	[.60, .93]	43	.88	.49	7	22	21	1	.24	.95	.55
NWF1	.90	[.77, 1.00]	16	.88	.95	7	2	41	1	.78	.98	.94
WIF1	.86	[.73, .98]	10	.88	.74	7	11	32	1	.39	.97	.76

Note. ORF = Oral Reading Fluency; AUC = Area under the curve; TP = True positive; FP = False positive; TN = True negative; FN = False negative; PPP = Positive predictive power; NPP = Negative predictive power; B/EI = Beginning/Early Intermediate; PSF = Phoneme Segmentation Fluency; DST = Developmental Spelling Test; NWF = Nonsense Word Fluency; WIF = Word Identification Fluency; I/EA = Intermediate/Early Advanced.

*Cut scores were identified that resulted in sensitivity levels as close to .90 as possible.

Table 12

Predictive Accuracy Indices by ELL Group for Winter Screeners on Below 25th Percentile of Spring ORF

Measure	AUC	95% CI	Cut Score*	Sn	Sp	TP	FP	TN	FN	PPP	NPP	Hit rate
B/EI (*n* = 55)												
PSF2	.61	[.45, .78]	62	.93	.23	14	31	9	1	.31	.90	.42
DST2	.83	[.71, 96]	53	.93	.55	14	18	22	1	.44	.96	.65
NWF2	.81	[.67, .95]	40	.93	.55	14	18	22	1	.44	.96	.65
WIF2	.99	[.97, 1.00]	17	.93	.95	14	2	38	1	.88	.97	.95
I/EA (*n* = 51)												
PSF2	.55	[.31, .78]	61	.88	.16	7	36	7	1	.16	.88	.27
DST2	.83	[.70, .97]	53	.88	.56	7	19	24	1	.27	.96	.61
NWF2	.83	[.72, .94]	40	.88	.77	7	10	33	1	.41	.97	.78
WIF2	.96	[.92, 1.00]	21	.88	.91	7	4	39	1	.64	.98	.90

Note. ORF = Oral Reading Fluency; AUC = Area under the curve; TP = True positive; FP = False positive; TN = True negative; FN = False negative; PPP = Positive predictive power; NPP = Negative predictive power; B/EI = Beginning/Early Intermediate; PSF = Phoneme Segmentation Fluency; DST = Developmental Spelling Test; NWF = Nonsense Word Fluency; WIF = Word Identification Fluency; I/EA = Intermediate/Early Advanced.

*Cut scores will be identified that result in sensitivity levels as close to .90 as possible.

Table 13

Identified Cut Scores by Group

Measure	Total	B/EI	I/EA
Fall			
PSF1	49	49	50
DST1	42	42	43
NWF1	34	30	16
WIF1	10	9	10
Winter			
PSF2	60	62	61
DST2	52	53	53
NWF2	49	40	40
WIF2	19	17	21

Note. B/EI = Beginning/Early Intermediate; I/EA = Intermediate/Early Advanced; PSF = Phoneme Segmentation Fluency; DST = Developmental Spelling Test; NWF = Nonsense Word Fluency; WIF = Word Identification Fluency.

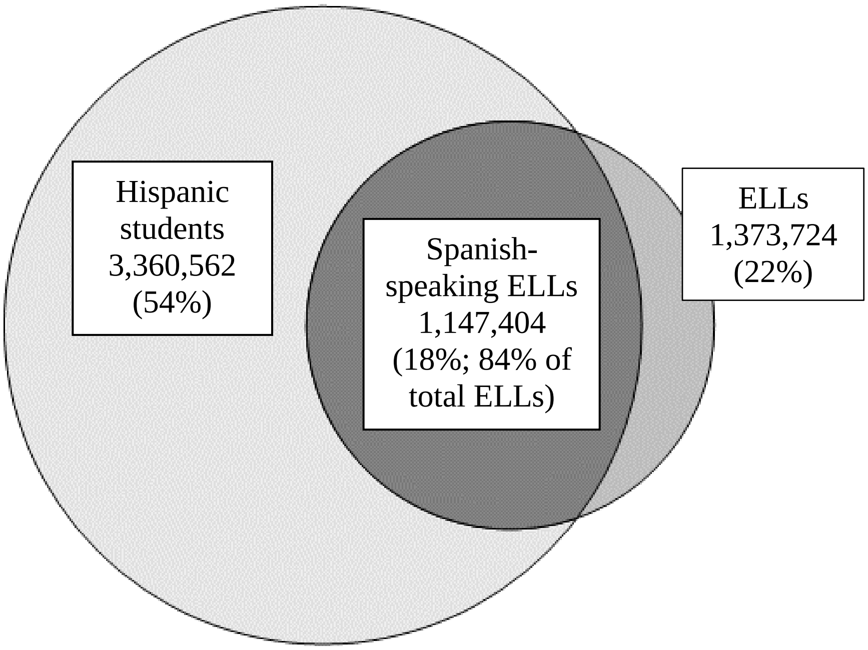
Hispanic students 3,360,562 (54%)
Spanish-speaking ELLs 1,147,404 (18%; 84% of total ELLs)
ELLs 1,373,724 (22%)

Figure 1. Hispanic students and English language learners (ELLs) as a percentage of California’s total K-12 population in 2015-2016 with Spanish-speaking ELLs identified as the overlap between the two.