An Evaluation of the Use of Oral Reading Fluency
As a Screening Tool
With Emerging Biliterates

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Kirsten M. W. Newell

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Robin S. Codding
Amanda L. Sullivan

July 2018

Acknowledgements

I would like to acknowledge the time and support provided by my dissertation committee members: Drs. Robin Codding, Amanda Sullivan, Kristen McMaster, and Tara Fortune. In particular, I appreciate Dr. Codding's suggestions and advice throughout the research process. I also appreciate the generosity of Dr. Fortune as she shared her expertise in immersion education and biliteracy acquisition. Finally, I would like to thank my school research partners Michael Mullins, Dr. Gesa Zinn, and Dr. Carol Mecklenburg for facilitating data collection and providing insight into the practical implications of this work.

Dedication

This dissertation is dedicated to my PhD pit crew.

Abstract

Students learning to read in more than one language, or emerging biliterates, are becoming increasingly common in schools. Early screening and identification of reading difficulties may lead to better outcomes for emerging biliterates as well as monolingual English students. Oral reading fluency (ORF) is one tool shown to be both a reliable measure of reading and an accurate method of identification of students at risk for poor reading outcomes. This project sought to build validity evidence for the use of ORF as a screening tool with emerging biliterates. Study one, a systematic review of literature, sought to synthesize available validity evidence for ORF with emerging biliterates. Studies were included that were empirical investigations about the use of ORF with emerging biliterate students in grades K through 8. All included studies (*n* = 25) were conducted with English language learners. Results suggested that although ORF is correlated with reading outcomes, the accuracy of ORF to identify emerging biliterates at risk of poor reading outcomes did not meet criteria. The strength of validity evidence differed by language proficiency of participants. Finally, there were substantial flaws in the quality of the reviewed studies. Study two, an empirical study at a German immersion school, sought to evaluate the use of ORF as a reading screening tool across German and English for students in third (*n* = 60), fourth (*n* = 60), and fifth grade (*n* = 42). Students were given ORF in English and German in winter, and a German proficiency and English reading exam in the spring. ORF in English, the first language of participants, was a good measure of reading and an accurate screening tool when predicting English and German outcome measures. English ORF outperformed German ORF as a predictor in all instances. Overall, results of the present studies suggest that though ORF has promise as a screening tool, more evidence is needed before it can be considered a valid, accurate screening tool for emerging biliterates.

Table of Contents

List of Tables

List of Figures

CHAPTER 1: Introduction

Response to intervention (RtI) is a framework that is predicated on data-based decision making and tiered systems of reading support for students, classrooms, and schools (Fuchs, Mock, Morgan, & Young, 2003; Fuchs & Young, 2012). This framework was designed as a prevention and early intervention model to improve the reading outcomes of all children. RtI is at least partially adopted in an estimated 48% of districts nationwide, and over 90% of these districts adopted RtI for reading (Spectrum K12, 2010). Preliminary outcomes suggest that RtI has resulted in decreases in referrals to special education as well as decreases in disproportionality (VanDerHeyden, Witt, & Gilbertson, 2007).

A central aspect of RtI is that school personnel can identify students struggling with reading at the start of the academic year and provide appropriate reading instructional interventions. Universal screening, or assessment of all students on selected skills (Jenkins, Hudson, & Johnson, 2007), is administered to identify those students who are at-risk for reading difficulties. Oral reading fluency (ORF) is a brief and easy universal screening tool that has a plethora of empirical support for this purpose (Fuchs, 2004; Kilgus et al., 2014). However, with increasing numbers of English language learners (National Center for Education Statistics; NCES, 2016) and dual language programs (Center for Applied Linguistics; CAL, 2011), it is necessary to evaluate the extent to which ORF retains its properties as a valid reading screening tool for linguistically diverse populations (Hosp, Hosp, & Dole, 2011). Therefore, the purpose of this project is to first synthesize available validity evidence about the use of ORF as a reading screening tool with any students who are learning to read in more than one language. Next, the project uses results of the synthesis to inform and conduct an empirical study that seeks to gather validity evidence about the use of ORF as a screening tool with students in an immersion setting, one subpopulation of students learning to read in more than one language.

Universal Screening of Reading

Universal screening consists of measuring student performance on selected skills to identify students who are at risk for not meeting standards (Jenkins, Hudson, & Johnson, 2007). That is, universal screening results identify those students not responding to the instruction they have been provided, so educators may adjust instruction or intervention to meet the needs of those students (Kilgus et al., 2014). Within the RtI framework this is a key point in accurate resource allocation (Jenkins et al., 2007).

One of the most important steps in universal screening is assessment selection. Ideal screening measures are efficient, to allow the feasible assessment of the whole student population (Deno, 1985; Jenkins et al., 2007). Screening measures must also be valid. Validation requires evidence to suggest that the inferences examiners make are accurate for the proposed interpretations and uses (Kane, 2013). An inference can be conceptualized as a leap in reasoning based on validity evidence. In the context of universal screening, criterion-related validity evidence is necessary to guarantee the *interpretation* inference that a reading screening tool truly measures reading skill (Kilgus et al., 2014). However, diagnostic accuracy evidence is also necessary to support the *use* inference that a measure can accurately classify or diagnose students at risk (Kilgus et al., 2014). Finally, validity evidence to support the use of an assessment with a specific population is also necessary (AERA, 2014). That is, validity evidence must be gathered with a population representative of the intended population for the assessment.

There are many types of reading assessments that might be used for universal screening. Informal inventories require students to read aloud and answer comprehension questions while the examiner rates students' performance (Parker et al., 2014). Computer-adaptive tests require students to answer a variety of questions within a software system. Results are usually reported in some type of standardized, norm-referenced score, across reading domains (cf. MAP by NWEA, 2013; aReading by TJCC, 2014; Clemens et al., 2015). Maze requires students to silently read a passage and select the correct missing word every seventh word (Graney et al., 2007). Word identification (Word ID) tasks require students to read word lists aloud while the examiner counts the number of words read correctly per minute (Deno, Mirkin, & Chiang, 1982).

ORF may be the most well-researched of these assessments, meeting the desirable screening assessment criteria of efficiency and validity (Deno, 1985; Jenkins et al., 2007). ORF requires a student to read a passage aloud as an examiner counts the words read correctly per minute. Therefore, in comparison to informal inventories, ORF is relatively efficient to administer. The use of ORF as a screener in reading is also supported by a larger body of criterion-related validity evidence (cf. Reschly et al., 2009) than other potential screening tools such as maze (Graney et al., 2007; Jenkins et al., 2007; Wayman et al., 2007) or computer adaptive tests (Clemens et al., 2015). For example, the number of words read aloud in one minute correlates both with standardized, norm-referenced tests of reading achievement as well as state accountability tests (Reschly et al., 2009; Wayman et al., 2007; Yeo, 2010), which suggests that ORF is truly a measure of reading.

Furthermore, there is some evidence to support the use of ORF to accurately classify or diagnose students at risk for not meeting reading standards (Kilgus et al., 2014), in contrast to other potential reading screening assessments such as informal inventories (Parker et al., 2014).

Despite the promise of ORF as an efficient and valid screening tool for English-speaking students, validity evidence must be gathered with a population representative of the intended population for the assessment. Validity evidence is sparse for students from diverse linguistic backgrounds (Sandberg & Reschly, 2011).

Bilingual populations. Bilinguals can be defined as individuals with sufficient language proficiency in more than one language, to "…function in a situation that is defined by specific cognitive and linguistic demands, to a level of performance indicated by either objective criteria or normative standards" (Bialystok, 2001, p. 18). This flexible definition includes English language learners (ELLs) in the United States, who speak a minority language at home, such as Spanish or Mandarin. This definition also includes majority language speakers learning a second language at school, such as English-speaking students in a Spanish immersion program.

Bilinguals can be described in relation to their first language (L1), their second language (L2), and their proficiency in each of those languages. Language proficiency can be used to specify different types or sub-populations of bilinguals. For example, ELLs may be further categorized as having "beginner" or "intermediate" English proficiency based on proficiency ratings or assessments (cf. Vanderwood, Tung & Checca, 2014).

Of particular interest in the present project is a population of school-age bilinguals learning to read in more than one language. These emerging biliterates display some level of language proficiency across more than one language and are simultaneously building reading skills in one or both language. Theories of biliteracy suggest that language proficiency and literacy skills across both languages, as well as similarity of languages learned, explain variability in reading skills for emerging biliterates (Bernhardt, 2005).

By far the largest group of bilinguals who are also emerging biliterates in K-12 public education is English language learners (ELLs). The percentage of k-12 students in the United States who are ELLs has steadily increased, from about 8.8 percent of students in the 2003-2004 school year (or 4.2 million) to about 9.3 percent of students in the 2013-2014 school year (or 4.5 million; NCES, 2016). The 9.3 percent of students who are ELLs have varying English proficiency, first-language proficiency, and varied overlap between their first language and English. Therefore, at least 9.3 percent of students in schools require screening measures validated for use with these sub-populations of ELLs.

Evidence for the interpretation and use of ORF as a reading screening tool for the special population of ELLs is limited. A recent summary of ORF evidence with ELLs cited only nine studies, and these studies found mixed results (Sandberg & Reschly, 2011). For example, it is unclear whether ORF in English correlates as strongly to broad reading outcomes for ELLs, particularly as language proficiency is not often accounted for (Baker & Good, 1995). Furthermore, there is little evidence to support the use of ORF as a tool that accurately identifies ELLs at-risk in reading (Hosp, Hosp, & Dole, 2011; Sandberg & Reschly, 2011).

ELLs are but one group of emerging biliterates in United States schools. A growing group of English-speaking students is attending immersion programs, where the goal of instruction is to build language proficiency and literacy in both English as well as a second language. The United States has seen an increasing number of immersion programs, from approximately 250 programs nationwide in 2003 (CAL, 2011) to nearly 2000 programs nationwide in 2011 (T. Fortune, personal communication, January 2, 2018). This group of emerging biliterates also varies in language proficiency and language similarity. There are no known studies that investigate the use of ORF as a screening tool with any students in immersion programs.

To summarize, ELLs and other populations of bilingual students are part of growing numbers of emerging biliterates in our public schools. Universal screening tools allow for more efficient and accurate resource allocation (Jenkins et al., 2007). Because validity evidence must be gathered with representative populations, validity evidence must be built for the use of reading screening tools such as ORF for emerging biliterates (García, McKoon, & August, 2006).

Purpose

Although there is considerable empirical support for using ORF to accurately identify English-only students at-risk for reading difficulties, it is unclear to what extent ORF has been validated for use with emerging biliterates. Given the increase of emerging biliterate students in schools, it is necessary to understand the functioning of ORF as a universal screening tool with these different populations of students. This project addresses the gap in knowledge via two methods: systematic synthesis of evidence and empirical investigation.

The first study summarizes research on the use of ORF to screen emerging biliterates. All available evidence to date has been gathered with English Language Learners. This study seeks to answer the following research questions: What is the relationship between ORF scores and reading comprehension measures for emerging biliterates? What is the relationship between ORF scores and other reading outcome measures for emerging biliterates? Finally, to what degree does evidence support the use of ORF as a tool to accurately identify emerging biliterates at-risk for poor reading outcomes?

The second study incorporates results of the synthesis to inform an empirical investigation with a different, unexplored population of emerging biliterates: English-speaking students learning German in an immersion school. Study two evaluates the criterion-related validity evidence and the diagnostic accuracy validity evidence of ORF screening within and across languages in an immersion school. The study seeks to answer the following research questions: To what extent is measurement of ORF in English and German predictive of English reading success as measured by a state reading assessment? Do the predictive properties of German and English ORF change across grades three to five? To what extent does German language proficiency mediate the relationship between German ORF and reading success and the relationship between English ORF and reading success? Finally, this study seeks to evaluate the ability of German and English ORF to accurately identify students who are at risk in reading. Specifically, how accurately can either German ORF, English ORF, or a combination of the two predict which students will not meet proficiency on the state reading assessment?

CHAPTER 2: Study One

Over the past several decades educators have implemented frequent, systematic data gathering to improve the quality of decisions made about student progress. A variety of measurement tools was developed to support data-based decision-making processes. Curriculum-based measurement (CBM) is one such tool. CBM can inform several different educational decisions including (a) determining students' risk for academic difficulties; (b) monitoring student progress, and (c) informing educational practices (Fuchs, 2004). In recent years the value of CBM oral reading fluency (ORF) as a universal screener in reading, to systematically measure all students' reading, has been widely supported with empirical evidence (Kilgus et al., 2014; Reschly et al., 2009). However, it is unclear to what extent ORF can be used for students learning to read in two languages (emerging biliterates), including English Language learners (ELLs). This is an important consideration given that the percentage of bilinguals in schools in the United States has increased from 8.8 to 9.3 (NCES, 2016). If CBM ORF is going to continue to serve as a critical screening tool to determine students' risk for reading difficulties, it is necessary to consider the validity of this tool for use with bilingual students, particularly emerging biliterates: those students learning to read in two or more languages (AERA, 2014; Kane, 2013).

Curriculum Based Measurement of Reading (CBM-R)

Curriculum-based measurement (CBM) is a type of assessment tool designed to closely align with students' curriculum. It was originally developed to measure the progress of students receiving special education services (Deno, 1985). CBM was developed from behavioral and idiographic frameworks (Deno, 1990). That is, student skill was measured by counting observable behaviors, and student progress was evaluated by comparing those counts over time within each student. For example, to administer a CBM-R oral reading fluency (ORF) probe, an examiner listens to a student read aloud for one minute. The examiner notes the number of words read correctly per minute (WCPM) by the student. This is in stark contrast to the evaluation of students' performance on standardized, published norm-referenced achievement tests. These tests were not sufficiently sensitive to determine whether students had made progress over a short period of time and were therefore not useful in immediate evaluation of the efficacy of specialized instruction (Deno, 1990). Although CBM encompasses multiple academic domains, by far the most researched CBM measure is CBM of reading (CBM-R; Kilgus et al., 2014; Wayman et al., 2007).

Initially, oral reading fluency probes were created from texts in the student's curriculum, to measure what a student could produce in texts aligned with daily instruction (Deno, 2003). When conceptualized as a behavioral measure, psychometric properties of CBM-R were not a focus. Rather, the intra-individual, idiographic comparisons drove decision making. Over time, standardized, leveled probe sets were developed by a variety of publishers to increase standardization of the measurement (cf. DIBELS by Good & Kaminski, 2011; FAST by TJCC, 2014; AIMSweb by Pearson, 2012). This shift from a behavioral measure to a psychometrically developed assessment parallels the shift in common uses of CBM-R ORF from a progress monitoring tool for individual students to a universal screening tool (Christ, Van Norman, & Nelson, 2016). This screening use requires normative comparisons outside of an idiographic framework.

Therefore, the psychometric properties become essential to build validity evidence, or empirical support for CBM-R ORF's interpretations and uses (Kane, 2013).

ORF validity evidence. Historically, validity evidence for CBM has been divided into three stages: (a) evidence related to the static score or level, (b) evidence related to the slope or trend, and (c) evidence related to instructional utility (Fuchs, 2004). Careful reading of this organizational system reveals that these three levels sought to validate the use of CBM-R ORF as a progress monitoring measure. The present project seeks to evaluate CBM-R ORF as a screening measure for a special population of students. Therefore, required validity evidence includes (a) criterion-related validity evidence to ensure the *interpretation* that CBM-R ORF truly measures reading skill, and (b) diagnostic validity evidence to support the *use* of CBM-R ORF to classify or diagnose students at risk (Kilgus et al., 2014). Since there is a large volume of research on CBM-R ORF, there have been multiple qualitative and quantitative summaries of the validity evidence supporting the use of CBM-R ORF. Results of these syntheses, as well as some more recent studies, are summarized below by type of validity evidence.

ORF criterion validity evidence. Initially, criterion validity evidence was built through correlations between ORF performance and standardized, nationally norm-referenced measures of reading. As interest in state accountability measures increased, ORF was correlated to end-of-year, state accountability tests as well (Reschly et al., 2009). Meta-analysis found moderate-to-large correlation (r = .67) between ORF WCPM scores and reading proficiency measures as a whole (Reschly et al., 2009). Higher correlations were found between individually administered reading tests (r = .83) than group administered reading tests (r = .71; Reschly et al., 2009). Based on meta-analytic results, researchers concluded that criterion validity met standards only if ORF and the outcome measure were given within the same academic year (Reschly et al., 2009; Yeo, 2010). Correlations between ORF WCPM scores and state accountability tests were also found to be moderate-to-large ($r = .65$; $r = .69$) across two separate meta-analyses (Reschly et al., 2009; Yeo, 2010). Researchers provided two hypotheses to explain the lower correlations between ORF and state accountability tests: (a) state accountability tests have lower quality of development than norm-referenced reading measures, and (b) state accountability tests align with specific state standards rather than broad reading achievement (Reschly et al., 2009; Yeo, 2010).

One long-standing concern in the validity of ORF as a screening tool is its ability to predict reading comprehension as well as decoding and fluency (Deno, 2003; Reschly et al., 2009; Wayman et al., 2007). In other words, educators are often concerned that ORF lacks validity to predict general reading outcomes, such as vocabulary and comprehension (Deno, 2003). Educators may describe students as fluent readers, but poor comprehenders; these students are also known as word callers (Hamilton & Shinn, 2003; Meisinger et al., 2009; 2010). To that end, several reviews have compiled evidence addressing this issue.

Numerous studies, synthesized qualitatively and quantitatively, determined that ORF scores have large, positive correlations with standardized, norm referenced reading achievement tests as well as other comprehension measures like CBM maze (Deno, 2003; Fuchs, 2004; Reschly et al., 2009; Wayman et al., 2007). This relationship between ORF and comprehension holds true even when speed of processing is used as a control variable (Kranzler, Brownell, & Miller, 1998). ORF scores were also found to correlate with measures of vocabulary and decoding, supporting the use of ORF as a measure of general reading skill (Reschly et al., 2009).

It is important to note that the relationship between ORF and general reading outcomes may change across grades. Wayman and colleagues (2007) summarized a study that found similar correlations of ORF and measures of comprehension across grades, of .82, .88, and .82 in grades two, three, and four respectively (Hosp & Fuchs, 2005), as well as a study that suggested different relationships among measures across third and fifth grades (Shinn et al., 1992). A cross-sectional study with a large sample size (*n*=5,472) found decreasing correlations of ORF and state reading test scores across grades three, five, and eight (Silberglitt et al., 2006). At least one study (Denton et al., 2011) has found weaker correlations between ORF scores and state accountability reading comprehension test in grades six, seven, and eight than reported in previous studies for earlier grades (cf. Reschly et al., 2009). Oral reading fluency was found to have a different predictive relationship with comprehension across grades four through eight (Yovanoff et al., 2005). When comparing ORF scores to state achievement tests, a multi-level analysis of studies across grades one through eight suggested that the relationship may be moderated by grade level (Yeo, 2010). Although there is evidence to suggest that fluency remains an important part of reading, the nature and strength of this relationship likely changes across grades, particularly as students enter middle school around sixth grade (Denton et al., 2011; Silberglitt et al., 2006; Yovanoff et al., 2005). Furthermore, there is very little evidence available pertaining to students in grades nine through twelve.

To summarize, ORF has been shown to be a robust indicator of reading performance for monolingual students, as measured by ORF scores' relationship to standardized, norm-referenced reading assessments and state accountability tests in reading. There remains some uncertainty about the stability or consistency of this relationship across higher grades. The three most comprehensive reviews of ORF criterion-related validity evidence included an average of 40 studies each (Reschly et al., 2009; Wayman et al., 2007; Yeo, 2010). Despite the large number of included studies, only three studies with ELLs (and none with other bilingual student populations) were included across all three syntheses (Baker & Good, 1995; Graves et al., 2005; Wiley & Deno, 2005). Even a review focused solely on the use of ORF with ELLs included only nine studies (Sandberg & Reschly, 2011). Furthermore, results were mixed about the strength of the relationship between ORF and any outcome measure for this population of students. For example, two studies cited in the synthesis by Wayman and colleagues (2007) suggested moderate to strong criterion-related validity for ELLs (Baker & Good, 1995; Wiley & Deno, 2005). Yet a multi-level analysis of the relationship between ORF scores and state achievement tests suggested that the strength of the validity coefficients is negatively affected by a greater proportion of ELLs in the sample (Yeo, 2010). This gap in the literature presents a severe limitation for the utility of ORF as a universal screener, particularly since validity evidence must be developed with a sample of students that is representative of the population to be assessed (AERA).

ORF diagnostic accuracy validity evidence. Diagnostic accuracy is an instrument's ability to correctly identify or diagnose a subpopulation of interest (Swets, Dawes, & Monahan, 2000). An instrument must be able to (a) predict or identify those at risk (sensitivity), as well as (b) correctly predict or identify those not at risk (specificity). Figure 1 illustrates the concept of diagnostic accuracy. There are several values calculated in diagnostic accuracy analyses. Sensitivity (SE) is the proportion of students who failed the outcome measure (true positive + false negative) who were correctly identified by the screening measure (true positive). Specificity (SP) is the proportion of students who passed the outcome measure (false positive + true negative) who were correctly identified by the screening measure (true negative). Ideally for educational decisions, SE and SP are balanced, with SE values about .8 and SP values above .7 (Kilgus et al., 2014; Silberglitt & Hintze, 2005; Swets, Dawes, & Monahan, 2000). This allows educators to be confident that they are neither over- nor under-identifying individuals with a screening tool. Two other important values can be calculated related to diagnostic accuracy: positive predictive power (PPP) and negative predictive power (NPP). PPP represents the proportion of students identified as at-risk on the screening measure (true positive + false positive) who actually failed the outcome measure (true positive). NPP represents the proportion of students who were identified as not at risk on the screening measure (false negative + true negative) who actually passed the outcome measure (true negative). Finally, there is overall accuracy of the measure. This is frequently calculated via Receiver Operating Curve (ROC) analyses and is often referred to as Area Under the Curve (AUC).

		Outcome Assessment Results	
		Fail	Pass
Screening Assessment Indicates	At Risk	True Positive	False Positive
	Not At Risk	False Negative	True Negative

Figure 1. Illustration of diagnostic accuracy, comparing a screening measure's results to the results of the outcome measure of interest.

To use ORF as a diagnostic screening measure, cut scores, or benchmarks, are developed. Students above a certain score are categorized as not at risk for our outcome of interest, and students below that score are categorized as at risk for our outcome of interest. Published tools often provide these benchmarks for grade and season (e.g. second grade fall, third grade spring, etc.; Good & Kaminski, 2011). Educators base resource allocation decisions on these benchmarks. For example, educators might provide an intervention to all students who are below benchmark, on the assumption that these students would otherwise not perform well on the outcome measure.

To evaluate the validity evidence in support of ORF interpretation via diagnostic accuracy, Kilgus and colleagues (2014) conducted what they called a diagnostic test accuracy meta-analysis, or DTAM. The researchers were interested in the quality of cut scores in ORF studies, whether cut scores varied across populations, and the overall diagnostic accuracy of ORF. Results compiled from 34 studies with students in grades 1 through 8 suggested that ORF can meet criteria for SE and SP as long as (a) screening tools and outcome measures are given within 12 months of each other, (b) local cut scores are developed, rather than generic publisher-provided cut scores based on national samples. Unfortunately, no studies with bilingual students were included in the analysis by Kilgus and colleagues.

Possible mechanisms of ORF. A large body of evidence suggests that ORF correlates with comprehension and broad reading outcomes. These correlations may change across time, or for special populations. To better understand why ORF may function differently for different populations or ages of students, it is helpful to examine the possible mechanisms of ORF within reading theory.

At first glance ORF is a simple measure. However, students learning to read must have integrated several complex, hierarchical skills to successfully read with accuracy, prosody, and speed (Ehri, 2005; Fuchs, Fuchs, Hosp, & Jenkins, 2001; NRP, 2000). Students must first understand and be able to manipulate phonemes, the smallest sound-units in a language. Next, students must link those sounds to letters. Then, students begin to sound out or decode words. As students build accuracy with decoding, reading speed increases. Finally, students achieve reading fluency, which incorporates accuracy, inflection or tone (prosody) and speed (NRP, 2000).

One hypothesis of why ORF scores correlate well with reading relates to efficiency of cognitive functioning (LaBerge & Samuels, 1974). When students are still developing decoding accuracy they direct cognitive resources, or attention, to reading each word. Once accuracy is achieved, decoding or reading a word takes fewer cognitive resources. Therefore, students have achieved automaticity with that skill; it can happen automatically, without direct attention. Similarly, in the so-called "simple view of reading", decoding x comprehension = reading (Gough & Tunmer, 1986).

As some researchers have pointed out, the automaticity model erroneously assumes that reading comprehension only occurs once the mechanics or decoding are sufficiently automatized (Fuchs, Fuchs, Hosp, & Jenkins, 2001). Evidence suggests that students use contextual clues, or using the meaning of text to help decode or read words, before they fully automatize decoding (Stanovich & Stanovich, 1995, as cited in Fuchs et al., 2001). In fact, it is possible that students have semantic networks that are automatically activated as they read, with this process happening automatically for good readers and with conscious attention for poor readers (Fuchs et al., 2001). In other words, good readers automatically use their knowledge of how words relate to each other in language for context, whereas poor readers use context clues with attention. As with the decoding automaticity theory (LaBerge & Samuels, 1974) and the simple view of reading (Gough & Tunmer, 1986), we arrive at the same conclusion: good readers read faster, poor readers read slower.

If automatic semantic network activation and the resultant comprehension is part of reading fluently, and ELLs have different comprehension in English than monolingual English students, then ORF WRCPM scores for ELLs may have a different relationship to comprehension measures or broad reading measures. Therefore, to better understand the functioning of ORF as a screening tool with biliterate students, it is necessary to explore some basic information about literacy acquisition in more than one language.

Biliteracy

The second language acquisition process can be a lengthy one. According to a study conducted with English language learners receiving English support in California schools, it can take three to five years for a language learner to acquire basic communication skills in face-to-face conversations, and five to seven years for a language learner to develop academic communication skills (Hakuta, Butler, & Witt, 2000). These two types of communication are often referred to as basic interpersonal communicative skills (BICS) and cognitive academic language proficiency (CALP; Cummins, 1980). However, it is important to note that language acquisition frameworks are not a replacement for biliteracy acquisition frameworks. Language proficiency is just one variable in a comprehensive model of second-language reading (Bernhardt, 2005; 2011).

Similar to monolingual literacy acquisition, students acquiring biliteracy must acquire multiple hierarchical skills at the word level and the text level in order to become accurate, fluent readers and comprehend text. There is some evidence that the simple view of reading also applies to students learning to read in a second language (Erdos et al., 2010). However, to paraphrase Grosjean (2010), a biliterate individual is not simply two literate individuals in one brain. That is, literacy skills in one language are not independent of literacy skills in another language. Consequently, theories of reading development in monolinguals may or may not transfer to the study of biliterate individuals (Bernhardt 2005; 2011), and the study of biliteracy acquisition involves more variables than the study of monolingual literacy acquisition.

Variables relevant to biliteracy. One comprehensive model of biliteracy suggests that variability in reading comprehension in a second language can be described primarily by knowledge of the second language (approximately 30% of variability explained) and reading in the first language (approximately 14% of variability explained; Bernhardt, 2005). These two broad areas of first language literacy and second language knowledge include multiple variables important to the exploration of the development of biliteracy: second-language oral language skills, concepts of print, metalinguistic awareness, and cultural variables (Bernhardt 2005; 2011; Bialystok, 2007).

Oral language. Oral language can be defined as spoken communication, including vocabulary, in a given language. Oral language could be described as part of the comprehension piece of the simple view of reading. As previously mentioned, bilinguals are rarely proficient at similar levels across languages. Research on bilingualism suggests that bilinguals typically have lower oral vocabulary in both languages than monolinguals have in one (Bialystok, 2007). Yet the results of multiple studies suggest that oral language proficiency plays an important role in second-language reading proficiency. For example, a study conducted with English-speaking and English-French bilingual students in a French immersion program investigated whether oral language predicted variability in reading outcomes after taking into account phonemic awareness (Erdos et al., 2010). Specifically, performance on a French vocabulary task given in Kindergarten was a significant predictor of French word decoding and reading comprehension in 1st grade. In a follow-up longitudinal study, the Kindergarten French vocabulary task was still a significant predictor of word reading in French in third grade (Genesee, Savage, Erdos, & Haigh, 2013).

Concepts of print. Concepts of print can be defined as an understanding of how written symbols are typically used to represent oral language. For example, in many languages print is read from the top left of a page toward the right, continuing down the page once the right side of one line of print is reached. In some languages phonemes, or sounds, are represented by symbols. In other languages symbols represent syllables or words. For English and other alphabetic languages, concepts of print include the link between letters and sounds. In other words, concepts of print include the word-level, decoding portion of the simple view of reading. There is some initial evidence that pre-literate bilinguals have better knowledge of concepts of print than monolinguals (Bialystok, 2007). Further research suggests that the similarity of written language across L1 and L2 can impact the extent to which literacy skills can transfer across languages (Bialystok, Luk, & Kwan, 2005; Koda, 2002; Koda, 2008).

Metalinguistic awareness. Metalinguistic awareness can be defined as thinking about the parts of language. This includes phonemic awareness, or awareness of the sounds in language. Phonemic awareness is a key skill in the acquisition of sound-symbol correspondence and eventual decoding. Metalinguistic awareness also includes awareness of grammar and syntax, or awareness of the use of words to construct thoughts or statements. In general, individuals exposed to more than one language have better metalinguistic awareness (Bialystok, 1988; 2007).

Cultural variables. Finally, it is important to note that cultural similarities and differences can influence biliteracy (Bialystok, 2007). For example, stories or experiences that are different in the second language may affect accessibility of content and, therefore, comprehension for individuals.

Cross-language transfer of literacy skills. A key issue to biliteracy development is the extent to which literacy skills in one language can transfer to a second or subsequent language. A number of studies attempted to provide evidence in support of the consistent cross-language transfer of literacy skills. For example, one seminal longitudinal study found not only that phonological awareness was strongly related to reading achievement within French and within English for French immersion learners, but that phonological awareness *across* languages was associated with decoding in French (Comeau et al., 1999). Results of a more recent study suggested that students' phonological awareness in Spanish was predictive of later word ID, with similar correlations to Spanish and English word ID lists (Lindsey, Manis & Bailey 2003).

Written language systems may play a role in cross-language transfer of literacy skills (Bialystok, Luk, & Kwan, 2005; Koda, 2002; Koda, 2008). For example, multiple-language learners did not transfer phonological awareness skills as readily from Chinese to English as they did from Spanish to English (Pasquarella et al., 2014). The researchers hypothesized that different phonics and symbols interfered with cross-language transfer from Chinese to English. Whereas English uses an alphabetic system in which symbols represent phonemes, Chinese uses a logographic system in which symbols represent words or word parts (for a complete explanation, see Koda, 2002). Furthermore, transfer may or may not require basic proficiency in a relatively similar first language before skills transfer to the second language (Koda, 2008), or basic proficiency in the second language before first-language skills can transfer (Lee & Schallert, 1997).

There is evidence that students may benefit from transferring phonological skills from an orthographically shallow language into a more complex language. Orthographically shallow languages have clear and consistent phoneme-symbol correspondence, as well as consistent spelling patterns and rules. Portuguese, Spanish, and German are relatively orthographically shallow languages. In contrast, English is a more complex language, with irregular spelling patterns and rules. A group of bilingual students (Portuguese/English) with learning disabilities outperformed monolingual students (English) on pseudoword reading and spelling tasks (DaFountoura and Siegel, 1995). Researchers hypothesized that cross-language transfer of literacy skills helped students perform in a more complex language.

There is also some evidence that certain aspects of language might not transfer as readily. For example, using factor analysis of longitudinal reading development of Turkish students learning Dutch, researchers found little relationship between grammatical or lexical skills across languages (Verhoeven, 1994). Phonological skills, however, did demonstrate strong cross-language relationships in these study results. Similarly, ELLs in the US demonstrated similar growth in decoding or word recognition to monolingual peers but fell significantly behind monolingual peers when vocabulary and comprehension were measured (Lesaux & Geva, 2006; Mancilla-Martinez & Lesaux, 2011). Results from a longitudinal study with French immersion students from Kindergarten through third grade indicated that phonological awareness in English correlated with future reading ability in French, but English vocabulary did not (Jared et al., 2011). There is some evidence that of a group of ELLs in elementary school, better readers utilized better reading comprehension strategies including self-monitoring, questioning, and making inferences (Durgunoglu, 2002). However, evidence on the cross-language transfer of comprehension skills such as self-monitoring or metacognition is not as robust, particularly for younger students as compared to adult second-language readers (Bernhardt, 2011). Therefore, it seems that phonological, word-level skills may transfer for young readers more readily than vocabulary or text-level skills.

Although there is widespread agreement that literacy in a first language affects literacy in a second language (Bernhardt 2005, 2011; Comeau et al., 1999; Koda, 2008), evidence on the transfer of literacy-related skills from one language to the next is more complex than Cummins' interdependence hypothesis might suggest. More recent theories of dual-language reading acquisition suggest compensatory processes (Bernhardt, 2005; 2011). That is, literacy skills in the first language serve to compensate for emerging skills in the second language, dependent on context and a large number of other variables (Bernhardt, 2005; Bialystok, 2007).

ORF Validity Evidence and Biliteracy

Despite the complexity of second-language literacy acquisition, some patterns emerge from the literature. First, literacy in the first language consistently explains a large amount of variability in literacy in the second language (Bernhardt, 2005). Furthermore, for orthographically similar languages, phonemic awareness and decoding skills seem to transfer across languages (Bialystok, Luk, & Kwan, 2005; Comeau et al., 1999; DaFountoura and Siegel, 1995; Lindsey, Manis & Bailey 2003). Finally, vocabulary, syntax and semantic network activation, and the resulting comprehension may not as clearly transfer across languages (Jared et al., 2011; Lesaux & Geva, 2006; Mancilla-Martinez & Lesaux, 2011; Verhoeven, 1994). This evidence has implications for the use of ORF with bilingual and biliterate students. For example, ORF in a student's first language may be necessary to effectively predict reading performance in a second language. If a strong vocabulary base has not yet been constructed in English (L2), then ORF scores may not demonstrate the same relationship to reading outcome measures for ELLs as it does for fully English-proficient readers.

In fact, there is some initial evidence in support of a different relationship between ORF and reading for ELLs: different numbers of word callers, or students who can decode but not comprehend. Researchers found a higher proportion of word callers in a third-grade population of ELLs when compared to monolingual students (Knight-Teague, Vanderwood, & Knight, 2014). In contrast, the number of word callers in the fifth-grade population of ELLs was similar to the number of monolingual word callers. In this cross-sectional study it is possible that the third-grade ELLs in this study had acquired less English than the fifth-grade ELLs. The study sample is not sufficiently well described to know.

To summarize, biliterate students are in different stages of language acquisition and literacy acquisition within different languages and, as such, reflect multiple populations rather than one. To build validity evidence for the interpretation and use of ORF as a screening tool, it is necessary to sufficiently describe biliterate populations, as well as demonstrate a strong relationship between ORF WCPM scores and reading outcomes for these populations.

Purpose

The purpose of this study was to critically evaluate the validity of ORF as a screening tool for use with emerging biliterates. Specifically, what is the relationship between ORF scores and reading comprehension measures for emerging biliterates? What is the relationship between ORF scores and other reading outcome measures for emerging biliterates? Finally, to what degree does evidence support the use of ORF as a tool to accurately identify emerging biliterates at-risk for poor reading outcomes?

Methods

Selection of Studies

This review focused on empirical, peer-reviewed studies, dissertations or technical reports published in English between 1980 and 2017. Articles were selected via electronic data-base search, including EBSCO, PsycINFO, and Google Scholar. The database search included dissertations as well as published articles. An ancestry search was used, through review of references of identified articles to target any other relevant studies. Finally, a search for publisher technical reports or manuals was conducted on publisher websites (AIMSweb, DMG/DIBELS, FastBridge Learning, easyCBM) as well as the National Center on Response to Intervention website (www.rti4success.org). Figure 2 displays the search process in detail.

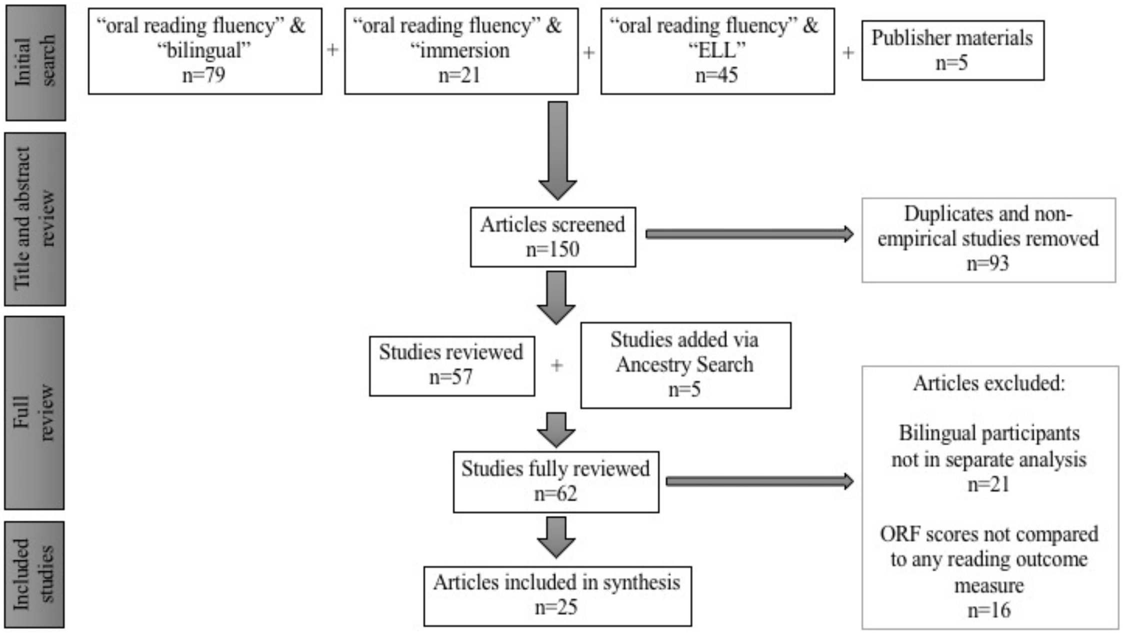
Initial search
"oral reading fluency" & "bilingual" n=79
+
"oral reading fluency" & "immersion n=21
+
"oral reading fluency" & "ELL" n=45
+
Publisher materials n=5
Title and abstract review
Articles screened n=150
Duplicates and non-empirical studies removed n=93
Full review
Studies reviewed n=57
+
Studies added via Ancestry Search n=5
Studies fully reviewed n=62
Articles excluded:
Bilingual participants not in separate analysis n=21
ORF scores not compared to any reading outcome measure n=16
Included studies
Articles included in synthesis n=25

Figure 2. Systematic search process overview.

Inclusion Criteria

Inclusion criteria narrowed the range of articles to only those that included bilingual students in Kindergarten through eighth grade, as the mechanisms of Oral Reading Fluency may change in grades nine through twelve. Studies were included only if the sample of bilingual students represented a separate population in final analyses and results. CBM oral reading fluency probes must have been administered in the standard manner with words read correct per minute scores calculated (cf. Shinn, 1989), which included (a) only words read correct in connected text, not word lists, and (b) passages leveled or controlled for difficulty in some manner. CBM ORF probes in any language were accepted. ORF scores must have been compared to some outcome measure to provide validity evidence. A final 25 studies were included in the synthesis.

Coding

Articles were coded based on CBM ORF probe language, CBM ORF probe developer or publisher, criterion measure type (comprehension, broad reading, state accountability test) and language, and language proficiency measure. Articles were also coded based on five different education types: (a) ELL support only, (b) developmental bilingual programs, in which language minority students learn in their first language as well as their second, (c) two-way immersion programs, where half of the student population speaks the minority language and half speaks the majority language, (d) heritage language immersion, such as students from indigenous cultures learning an endangered heritage language, and (e) foreign language immersion, in which students of a majority language are immersed in a foreign language, (Christian, 2011; Howard et al., 2007). Articles were also coded on characteristics of bilingual participants (first and second languages spoken, similarity of languages spoken, measure of proficiency in languages spoken, time of exposure to languages spoken). Finally, articles were coded based on the type of validity evidence they provide: correlational or diagnostic. Studies that provided a correlation between ORF and an outcome measure, or provided regression models to determine a correlation, were coded as correlational/interpretation evidence. Studies that provided the results of a diagnostic accuracy analysis (ROC or cross-tabulated accuracy) were coded as diagnostic/use evidence.

Table 1.

Rubric for quality indicators rating.

Quality Indicator	1 – Not met	2 – Partially met	3 - Met
Q1: Description of participants	Registration forms or other district-wide method for ID of ELLs used.	Language background of participants partially described, such as estimate of sample language proficiency	Participant language background individually assessed and reported. Languages spoken provided, proficiency in both languages measured
Q2: Description of dual-language instruction	Language instruction not specified, or described only as "ELL services"	Language instruction described with some time estimate (e.g. pull-out ELL services twice per week)	Language instruction described, with content of instruction and time provided (e.g. ELL services twice per week for 20 minutes each, focused on unknown vocabulary words from classroom
Q3: Description of criterion measure	Outcome measure poorly described	Some information provided, such as either reliability or validity	Validity evidence and other psychometric properties of the outcome measure provided, measure clearly aligns with research questions
Q4: Sample size and sampling strategy	No power analysis conducted, convenience sampling	Some attempt to sample randomly or address representativeness, appropriate sample size	Power analysis conducted before analysis, random sampling or sampling carefully controlled
Q5: Assumptions of analysis discussed and addressed	Assumptions neither discussed nor addressed	Assumptions mentioned but not fully addressed	Assumptions discussed and addressed
Q6: Descriptive statistics provided for each relevant variable	None	Some	All

To capture the quality of the literature, articles were also coded based on six indicators (Gersten et al., 2005; Jitendra, Burgess, & Gajria, 2011; Talbott, Maggin, VanAcker, & Kumm, 2017). The quality indicators are presented in Table 1. Quality indicator one (Q1) considers whether authors sufficiently described the participants, including participant language proficiency and languages spoken. Quality indicator two (Q2) rates the extent to which authors described the type of instruction received by participants. The third quality indicator (Q3) rates authors on whether outcome or criterion measures were sufficiently described. Quality indicator four (Q4) addresses whether authors considered sample size and sampling issues, and quality indicator five (Q5) reviews whether statistical analysis assumptions were considered. Finally, quality indicator six (Q6) addresses whether authors sufficiently describe included variables with descriptive statistics.

Data Analysis

First, numbers of studies in each category within all coded variables were calculated. Then, key results were organized in a table (Appendix A). Information was synthesized by measures used, participants, and types of validity evidence. Finally, quality indicators were aggregated for all studies.

Interrater reliability. A graduate student trained in research methodology coded approximately 20% (n=6) of the 25 included studies. After training on the rubric and definition of variables and quality indicators, agreement was 100% on all but three of the coded variables. One disagreement on the type of criterion measure (broad reading vs. comprehension) was resolved with discussion. Two of the quality indicators also had 83% agreement, or agreement on five of six studies. After discussion, 100% agreement was reached on these two variables, description of language proficiency (Q1) and description of instruction (Q2).

Results

Overall, 25 studies met inclusion criteria; 15 peer-reviewed, published articles, 9 dissertations, and 1 technical report. A comprehensive summary of included studies is presented in Appendix A. The majority of studies presented only correlational validity evidence (*n* = 18; 72%), and a few studies presented only diagnostic validity evidence (*n* = 3; 12%). Several studies presented both correlational and diagnostic validity evidence (*n* = 4; 16%).

Measures

Several types of measures were used in the reviewed studies. All studies included an ORF measure and a criterion measure. Some studies included a language proficiency measure (*n* = 11; 44%), and some studies included the measurement of other variables relevant to reading and language proficiency such as vocabulary or oral language (*n* = 10; 40%).

ORF measures. Many included studies (*n* = 12; 48%) used the ORF measure that is a part of Dynamic Indicators of Basic Early Literacy Skills (DIBELS). Three of those twelve studies (Baker, 2007; Baker, Park & Baker, 2012; Felt, 2015) also included a Spanish measure of ORF, Indicadores Dinámicos del Éxito en la Lectura – Fluidez in la Lectura Oral (IDEL FLO; Cummings, Baker, & Good, 2006). Four total studies (Baker, 2007; Baker, Park & Baker, 2012; Felt, 2015; Ganan, 2012) included ORF measures in any other language than English, all of these measured ORF in Spanish with IDEL FLO. Three studies (12%; Farmer, 2013; Quirk & Beem, 2012; Stokes, 2010) used AIMSweb ORF measure (Pearson, 2012), one technical report (Sáez et al., 2010) was from easyCBM, one study (Jimerson, Hong, Stage, & Gerber, 2013) that used the Oral

Reading Assessment Level by Jimerson (ORAL-J; Jimerson, 2000), and one study (Crosson & Lesaux, 2010) used a raw score from the passage reading fluency task on the Gray Oral Reading Test (GORT; Weiderholt & Bryant, 2001). The remaining six studies used reading passages pulled from the curriculum to measure ORF in a method in keeping with CBM ORF as defined by Shinn (1989).

Criterion measures. Appendix A includes criterion measures used in each study. All studies but two (8%; Baker, 2007; Baker, Park, & Baker, 2012) used criterion measures in English. Both studies with non-English criterion measures used the Aprenda Spanish reading comprehension task (Harcourt, 2005) as well as an English outcome measure.

Twelve studies (48%) measured reading comprehension. The most common measure used was the Stanford Achievement Test – 10th Edition reading comprehension task (SAT10; Harcourt Brace, 2003; *n* = 4; 16%). Two studies (8%; Crosson & Lesaux, 2010; Quirk & Beem, 2012) used the Gates Macginitie comprehension task (MacGinitie, MacGinitie, Maria, & Dreyer, 2000). Two studies (8%; Kim, 2012a; Nam, 2012) used the Woodcock Reading Mastery Test – Revised passage comprehension task (WRMT-R; Woodcock, 1987). One study (Grasparil & Hernandez, 2015) combined the California Achievement Test – Reading Comprehension – Sixth Edition (CAT6; CTB McGraw Hill, 2001), the reading comprehension portion of the California English Language Development Test (CELDT; California Department of Education, 2007), and the reading comprehension portion of the California Standards Test (CST; Educational Testing Service, 2007) into a comprehension composite. Finally, one study (Millett, 2011) used the Terra Nova comprehension task (McGraw-Hill, 2003), one study (Baker & Good,

1995) used the Stanford Diagnostic Reading Test comprehension task (SDRT; Karlsen & Gardner, 1985), and one study (Pretorius & Spaull, 2016) used a researcher-created reading comprehension task.

Many studies used available state assessment results (*n* = 13; 52%) as criterion measures. In general, state accountability assessments are designed to measure whether students have achieved state standards in reading and language arts domains. One study (Grasparil & Hernandez, 2015) used separate vocabulary and comprehension scores that are a part of the CST (Educational Testing Service, 2007). All other studies with state assessment results used the overall, broad reading score. Two studies (Muyskens, Betts, Lau, & Marston, 2009; Wiley & Deno, 2005) used the Minnesota state accountability test (Minnesota Comprehensive Assessment; MCA; Minnesota Department of Education, 2015). Two studies (Farmer, 2013; Ganan, 2012) used the Illinois state accountability test (Illinois State Achievement Test; ISAT; Illinois State Board of Education, 2010). Two studies (Baker, Park & Baker, 2012; Sáez et al., 2010) used the Oregon state accountability test (Oregon Assessment of Knowledge and Skills; OAKS; Oregon Department of Education, 2008), and two studies (Kim, 2012b; Vanderwood, Tung, & Checca, 2014) used the California state accountability test (CST; Educational Testing Services, 2007). One study (Stokes, 2010) used the Arizona state test (Arizona Instrument to Measure Standards; AIMS; Arizona Department of Education, 2009), one study (Hosp, Hosp, & Dole, 2011) used the Utah state test (Utah State Criterion-Referenced Tests; Utah State Office of Education, 2007) one study (Echols, 2010) used the Washington state test (Washington Assessment of Student Learning; WASL; Pearson, 2007), and one study (Felt, 2015) used the Wisconsin state test (Wisconsin Knowledge and Concepts Exam; WKCE; CTB McGraw-Hill, 2014.

Three studies (12%; Baker & Good, 1995; Betts, Muyskens, & Marston, 2006; Jimerson et al., 2013) used a broad reading measure not part of a state accountability test. These assessments were the SDRT (Karlsen & Gardner, 1985), the Northwest Achievement Levels Test (NALT; Northwest Evaluation Association, 2002), and the Stanford Achievement Test – Ninth Edition (SAT9; Harcourt, 1997).

Language proficiency measures. Eleven studies (44%) used measures described by publishers as measures of language proficiency. Five studies (20%; Grasparil & Hernandez, 2015; Kim, 2012b; Nam, 2012; Quirk & Beem, 2012; Vanderwood, Tung, & Checca, 2014) measured English language proficiency with the CELDT (California Department of Education 2007). This assessment includes listening, reading, speaking, and writing, and places students in one of five proficiency levels or categories: beginning, early intermediate, intermediate, early advanced, and advanced.

Two studies (8%; Baker & Good, 1995; Jimerson et al., 2013) measured English language proficiency with the Language Assessment Scale (LAS; DeAvila & Duncan, 1977). The assessment has tasks designed to measure vocabulary as well as receptive and expressive language. Scores range from 0 to 100, and a student with a score below 74 would be categorized as limited English proficient (Dalton, 1979)

Two studies (8%; Farmer, 2013; Felt, 2015) measured English language proficiency with the Accessing Comprehension and Communication in English State-to-State for English Language Learners (ACCESS for ELLs; WIDA Consortium, 2007) assessment. ACCESS assesses listening, speaking, reading and writing. These areas are combined to yield an overall score. Overall scores range from 1 (beginning English language proficiency) to 6 (the highest possible English language proficiency).

One study (Crosson & Lesaux, 2010) measured English listening comprehension with the Woodcock Language Proficiency Battery – Revised listening comprehension task (WLPBR; Woodcock, 1991). The entire WLPB-R includes oral language as well as reading and writing tasks. However, the only task used by a study in the present review requires students to provide a missing word within a passage that is read aloud.

Finally, one study (4%; Stokes, 2010) measured English language proficiency with the Arizona state English Language Learner Assessment (AZELLA; Arizona Department of Education, 2007). The assessment is designed to measure reading, writing, vocabulary, and oral communication. Students' English proficiency is categorized by the AZELLA as pre-emergent, emergent, basic, intermediate, or proficient.

Measures of other variables relevant to language and reading proficiency. Five studies (20%) included a measure of English vocabulary. Two of those studies (Johnson et al., 2009; Millet, 2011) used only the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 2007). One study (Riedel, 2007) used the Group Reading Assessment and Diagnostic Evaluation (GRADE; Williams, 2001) vocabulary task. One study (Kim, 2012a) measured English vocabulary with the PPVT as well as the vocabulary task from the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999), and also measured English oral proficiency with the Woodcock Johnson III oral proficiency task (WJ-III; Woodcock, McGrew, & Mather, 2001). One study (Crosson & Lesaux, 2010) measured Spanish vocabulary with the Test de Vocabulario en Imagenes

(TVIP; Dunn, Padilla, Lugo, & Dunn, 1986), as well as English vocabulary with the PPVT, and English listening comprehension with the WLPB-R listening comprehension task. Ten studies (40%) included only a measure of reading and a measure of ORF and did not include a measure of language proficiency or a measure of any other relevant variables.

Participants

The majority of studies included Spanish-speaking students learning English (*n*=14; 56%). One study included participants who spoke Korean as a first language (Nam, 2012), and one included participants whose first language was one of several South African languages (Pretorius & Spaull, 2016). The remaining studies (*n* = 9; 36%) did not disaggregate participants based on first language spoken, but simply specified that participants were ELLs. Nearly half of the studies provided no information about the English language proficiency of participants (*n* = 12; 48%). Only one study (Crosson & Lesaux, 2010) measured language proficiency in both languages used by participants. In those studies that described language proficiency of participants with categories, most studies had the majority of students at the intermediate English proficiency level (Farmer, 2013; Felt, 2015; Kim, 2012b; Nam, 2012; Quirk & Beem, 2012; Stokes, 2010). English proficiency tended to increase as grade levels increased (Millett, 2011).

Five studies (20%) described instruction only as “ELL.” Several studies indicated that English-only instruction was provided (*n* = 8; 32%). One study described instruction as foreign language immersion (Pretorius & Spaull, 2016), and one study described instruction as two-way immersion (Felt, 2015). Four studies (Baker, 2007; Baker, Park, & Baker, 2012; Crosson & Lesaux, 2010; Ganan, 2012) described the instruction as transitional or developmental bilingual. The remaining studies did not provide a description of instruction (*n* = 6; 24%).

Quality of Included Studies

Results of the quality indicator scoring are presented in Table 2. Most studies (*n* = 22; 88%) provided descriptive statistics for the relevant variables included. The description of the criterion measure used met quality criteria for fifteen studies (60%) and partially met criteria for nine studies (36%). Twelve studies (48%) failed to sufficiently describe assumptions of analyses conducted. Sample sizes and sampling strategies criteria were partially met for 13 studies (52%) and not met for 12 studies (48%).

The majority of studies did not meet criteria for the description of participants and the description of instruction. Only one study (Crosson & Lesaux, 2010) sufficiently described the language proficiency of participants in both languages. Six studies sufficiently described the language, time, and content of provided dual-language instruction.

Table 2.

Quality of studies included in synthesis.

Quality Indicator	Not met	Partially met	Met
Q1: Description of participants	12	12	1
Q2: Description of dual-language instruction	15	4	6
Q3: Description of criterion measure	1	9	15
Q4: Sample size and sampling strategy	12	13	0
Q5: Assumptions of analysis discussed and addressed	12	7	6
Q6: Descriptive statistics provided for each relevant variable	3	0	22

Validity Evidence for the Use of ORF with Emerging Biliterates

To address the research questions about criterion-related validity evidence and diagnostic accuracy validity evidence, the validity evidence is synthesized by type.

Criterion-related validity evidence. Criterion-related validity evidence is synthesized by criterion measure type (broad reading vs. reading comprehension) to address the following research questions: a) What is the relationship between ORF scores and reading comprehension measures for emerging biliterates? b) What is the relationship between ORF scores and other reading outcome measures for emerging biliterates?

Relationship between ORF and reading comprehension measures. Twelve studies (48%) included a reading comprehension criterion measure. The majority of these studies (*n* = 10) measured this relationship in English only. Across studies, the concurrent correlation between English ORF and English reading comprehension for ELLs ranged from .61 (*n* = 1772; Pretorius & Spaull, 2016) to .82 (*n* = 102; Nam, 2012).

Two studies (Baker, 2007; Baker, Park, & Baker, 2012) compared ORF and reading comprehension with Spanish as well as across English and Spanish. Within Spanish, the concurrent correlation between Spanish ORF and Spanish reading comprehension was .62 (*n* = 88; Baker, 2007). Across languages, the correlation between English ORF and Spanish reading comprehension was .54 (*n* = 88; Baker, 2007), and the correlation between Spanish ORF and English reading comprehension was .55 (*n* = 88; Baker, 2007).

One study measured the relationship between English reading comprehension and English ORF disaggregated by English language proficiency of ELLs (Nam, 2011). The overall concurrent correlation between ORF performance and reading comprehension was .82 for all participants. However, for participants with beginning English language proficiency, the correlation between ORF and reading comprehension was .87 (*n* = 27).

For participants with intermediate English language proficiency, the correlation between ORF and reading comprehension was .75 (*n* = 42). For participants with advanced English language proficiency, the correlation between ORF and reading comprehension was .69 (*n* = 33).

In six studies (20%), there was some evidence that the relationship between ORF and reading comprehension was affected by a third variable. For example, three studies (Garasparil & Hernandez, 2015; Millett, 2011; Riedel, 2007) found that English vocabulary better explained ELLs' English reading comprehension than English ORF did. Where available, correlations between vocabulary and comprehension measures are included in Appendix A. Another study's (Crosson & Lesaux, 2010) results suggested that students with low listening comprehension had low reading comprehension regardless of the ORF score. Finally, one study (Quirk & Beem, 2012) found a higher percentage of word callers (students with strong ORF scores but low reading comprehension) than previously found with monolingual English speakers.

Relationship between ORF scores and broad reading measures. Sixteen studies (64%) included a broad reading criterion measure. The majority of these studies (*n* = 14) measured this relationship in English only. Across studies, the concurrent correlation between English ORF and English broad reading measures for ELLs ranged from .36 (*n* = 109; Kim, 2012b) to .75 (*n* = 33; Sáez et al., 2010). The relationship between state accountability assessments and ORF ranged from .56 (*n* = 142; Baker, Park Baker, 2012) to .75 (*n* = 33; Sáez et al., 2010). The relationship between other standardized norm-referenced broad reading measures ranged from .53 (*n* = 26; Baker & Good, 1995) to .72 (*n* = 85; Jimerson et al., 2013).

Two studies (8%; Felt, 2015; Ganan, 2012) measured ORF and broad reading across languages. The correlation between Spanish ORF and a broad reading measure ranged from .48 (n = 234; Ganan, 2012) to .68 (n = 238; Felt, 2015).

Two studies measured the relationship between English ORF and English broad reading disaggregated by language proficiency level (Vanderwood, Tung, & Checca, 2014; Kim, 2012b). The correlations for students rated as beginners were .53 (*n* = 122; Kim, 2012b) and .66 (*n* = 198; Vanderwood, Tung, & Checca, 2014). The correlations for students rated as having intermediate English proficiency were .38 (*n* = 291; Kim, 2012b) and .52 (*n* = 193; Vanderwood, Tung, & Checca). Finally, the correlations for students rated as having advanced English proficiency were .36 (*n* = 109; Kim, 2012b) and .66 (*n* = 159; Vanderwood, Tung, & Checca).

Diagnostic accuracy validity evidence. Seven studies provided evidence related to diagnostic accuracy (28%). One study (Johnson, Jenkins, Petscher, & Catts, 2009) used a reading comprehension measure and the rest used state accountability assessments as outcome measures. Results of four studies (Echols, 2010; Famer, 2013; Hosp, Hosp, & Dole, 2011; Johnson et al.) included sensitivity estimates that met criterion, but in each of these studies but one the associated value of specificity did not meet criterion. Results of four studies (Ganan, 2012; Hosp, Hosp, & Dole; Vanderwood, Tung, & Checca, 2014; Muyskens et al., 2009) included specificity estimates that met criterion, but in each of these studies but one the associated sensitivity estimates did not meet criterion. The only diagnostic accuracy values that met criteria for overall accuracy, sensitivity, and specificity was from an analysis with a sample of first grade students (*n* = 403) from a larger study (Hosp, Hosp, & Dole).

One study (Vanderwood, Tung, & Checca, 2014) disaggregated diagnostic accuracy by language proficiency, and predicted spring broad reading success from winter ORF scores. When sensitivity was set to .90, the specificity values for students with beginning, intermediate, and advanced English language proficiency were .20, .30, and .50. For native English speakers, the specificity value was .55.

Discussion

The purpose of this systematic synthesis of evidence was to critically evaluate the validity evidence available in support of ORF as a screening tool with emerging biliterates. Specifically, the review sought to synthesize evidence related to the relationship between ORF scores and reading comprehension measures, as well as the relationship between ORF scores and broad reading measures. Finally, the review sought to evaluate the diagnostic accuracy validity evidence in support of the use of ORF as a tool to accurately identify emerging biliterates at-risk for poor reading outcomes.

The reviewed studies focused primarily on ELLs, many of whom spoke Spanish as a first language. Language proficiency, language background, and language instruction were generally not well described. Correlations between ORF scores and reading outcome measures were similar for ELLs as for monolingual English speakers. For example, the correlations between ORF and reading comprehension measures tended to decrease as grades increased for ELLs, similar to the pattern seen with monolingual English speakers. However, the diagnostic accuracy of ORF to identify ELLs as at-risk on a reading outcome was not as strong as the diagnostic accuracy of ORF to identify monolingual English speakers.

ORF as a Measure of Reading for Emerging Biliterates

In the reviewed studies, correlations between ORF scores and reading comprehension scores were higher in lower grades. The trend of decreasing correlations as grades increase is also seen with monolingual English-speaking students (Silberglitt et al., 2006). As previous researchers have suggested (Silberglitt et al., 2006), lower correlations in higher grades may be related to the diminishing importance of rate or fluency among predictors of reading outcomes. In the present synthesis, there is an added confound of grade and language proficiency. That is, ELLs may have better language proficiency in higher grades, and ELLs may receive different types of instruction and support across grades and levels of language proficiency. The results of several studies (Nam, 2012; Vanderwood, Tung, & Checca, 2014) included in the synthesis suggested that correlations between ORF and reading outcome measures are different for ELLs with different levels of English language proficiency. Therefore, it is unclear whether changes in correlations across grades are related to different relative importance of aspects of reading, English language proficiency, type of instruction, or a combination of all of these variables.

In the reviewed studies that used broad reading rather than reading comprehension measures, correlations between ORF and reading outcomes did not have a decreasing pattern across grades. One possible reason is the inclusion of state accountability assessments as broad reading measures. As suggested in previous reviews of ORF evidence (Reschly et al., 2009), state accountability tests are designed to measure state standards, which may result in measurement of slightly different skills than broad reading assessments such as the SAT10. Furthermore, some state accountability assessments may have a lower quality of development than other standardized, norm-referenced reading measures commonly given. Despite these differences, almost all of the correlations between ORF scores and broad reading measures within reviewed studies fall within the range of moderate to strong, which mirrors the results found with monolingual English-speaking students.

In the reviewed studies, correlations between ORF scores and reading comprehension scores were far stronger within English than within Spanish or across languages (English-Spanish or Spanish-English). These results suggest that unlike early literacy screeners that investigate phonological awareness (Erdos et al., 2010), ORF does not function equally well as a predictor within and across languages. These results support the concept that ORF measures more than just decoding or word-level tasks.

Quality of reviewed studies. To summarize overall synthesis results, the correlations between ORF and reading outcome measures for ELLs in the reviewed studies are similar to the correlations between ORF and reading outcome measures for monolingual students in previous research (Deno, 2003; Reschly et al., 2009; Wayman et al., 2007). However, there are several substantial limitations in the reviewed studies that prevent the interpretation of these correlations as evidence to support ORF as a reading measure for emerging biliterates. These limitations include failure to sufficiently describe the language and content of literacy instruction of participants, failure to describe the language background of the participants, and failure to include language proficiency and other variables relevant to L2 reading in analyses.

Language and content of literacy instruction. Many of the reviewed studies did not describe the literacy instruction provided to emerging biliterates. This is problematic because previous work on reading outcomes for emerging biliterates suggests that the trajectory of reading outcomes may vary based on the amount of literacy instruction provided in L1 as compared to L2. For example, ELLs who receive English-only instruction may acquire English language skills more quickly than students who receive dual-language instruction, but students in dual-language instruction may acquire literacy and English/Language Arts skills more quickly than students who receive English-only instruction (Umansky, Valentino, & Reardon, 2016). Furthermore, there is much evidence to suggest that instruction in phonemic awareness and phonics is as valuable for ELLs as it is for monolingual English students (August, McCardle, & Shanahan, 2008). Without clear description of content and language of instruction, it is difficult to generalize the results of the included studies to other populations of emerging bilinguals.

Description of language background. Across both reading comprehension and broad reading, many of the studies did not sufficiently describe the language background of the included participants. For example, several studies relied on extant data and the student demographic flag of "ELL," rather than any language proficiency information. The use of a demographic flag or label of ELL is problematic because different districts may have different criteria to identify ELLs. For example, in one sample, ELLs may be identified based on a specific cut score on a language proficiency measure (cf. Dalton, 1979). In another sample, ELLs may be identified based on a response on a parent form, without any measure of language proficiency. These two different samples may represent different populations, which would limit the comparability and generalizability of study results. Furthermore, participants identified solely on the basis of a label of ELL likely speak a variety of first languages. Previous research suggests that the similarity of first and second languages may influence the type and amount of reading skills that transfer across languages (Koda, 2008). Therefore, again, the label of ELL represents multiple different populations for whom the relationship between ORF and reading outcomes may be different.

Language proficiency is another aspect of language background not sufficiently described by many of the included studies. Based on Bernhardt's (2005) compensatory model of variables that contribute to L2 reading comprehension, language proficiency in L2 is an important variable. Higher levels of L2 language proficiency, according to the model, would be related to higher levels of reading comprehension if L1 reading skills were held constant. The simple view of reading within a language would suggest the same: increase in language comprehension (as measured by language proficiency variables) would yield increased reading comprehension outcomes. In other words, language proficiency differences may be related to differences in the relationship between ORF and reading outcomes, and therefore must be measured and included in studies.

Three studies disaggregated results based on English language proficiency. The pattern of correlations between ORF and outcome measures was different when ELLs were disaggregated by English language proficiency. The correlation between ORF scores and reading comprehension scores decreased as language proficiency increased (Nam, 2012). For ELLs in third grade, the correlation between ORF scores and state accountability assessment scores also decreased with increasing language proficiency (Kim, 2012b). For second grade ELLs, the correlation between ORF scores and state accountability assessment scores was weakest for students with intermediate language proficiency skills (Vanderwood, Tung, & Checca, 2014). In each of these studies, mean performance on ORF tasks as well as mean performance on reading outcome measures increased across increased levels of language proficiency. These results lend support to the necessity of disaggregation of validity evidence by language proficiency.

Finally, language background is important to describe in order to understand the relationship between ORF and reading outcomes at different grades for emerging biliterates. In several studies, results were disaggregated by grade but not by language proficiency. Descriptive statistics provided the language proficiency of participants in each grade. As grades increased, English language proficiency increased. In other words, changes in results by grade were confounded with changes in language proficiency. By measuring and including language proficiency in analyses, it would be possible to determine how ORF correlates with measure of reading for emerging biliterates across grades.

Measurement of variables relevant to L2 reading. Another theme that emerges from the reviewed evidence is a pattern of some other variable that might explain the relationship between ORF and reading for ELLs, or that might even be a better predictor of overall reading. This was measured as vocabulary (Riedel, 2007; Grasparil & Hernandez, 2015), language proficiency (Nam, 2011), or oral language (Crosson & Lesaux, 2010; Kim, 2012a). It is possible that for ELLs, ORF functions as a measure of the extent to which L1 literacy skills like phonological awareness have transferred across languages to support English decoding. In other words, based on Bernhardt's compensatory model of second language reading, ORF may measure L1 literacy and ignore L2 language knowledge. The model would then suggest that measurement of L2 knowledge would be necessary to more accurately predict reading comprehension. If for emerging biliterates ORF measures only decoding, or word-level reading skills, and does not sufficiently explain reading comprehension, we would see evidence of increased numbers of word callers among ELLs (Quirk & Beem, 2012). Two studies included in the present review found a higher percentage of word callers among ELLs than previously reported among monolingual English speakers (Felt, 2015; Quirk & Beem, 2012). To summarize, the results of the present synthesis suggest the need for future research to measure oral language, vocabulary, and/or language proficiency to build validity evidence for the use of ORF as a measure of reading outcomes.

ORF as a Screening Tool for Emerging Biliterates

In the current review, studies that provided diagnostic accuracy evidence did not all provide sufficient description of student language background or proficiency, nor language or type of instruction. It is important to interpret results of this synthesis with that limitation in mind. There was a clear pattern in the reviewed studies that suggested that accurate identification of students at risk of failing an outcome assessment would lead to over-identification of ELLs, or accurate identification of students passing an outcome assessment would lead to under-identification of ELLs. In other words, if sensitivity was maximized, specificity was too low, and if specificity was maximized, sensitivity was too low. These results lend support to the careful analysis of prediction outcomes over and above an overall accuracy, AUC, or "hit-rate" value.

Several studies provided evidence that suggested that cut scores for ELLs on screening measures need to be different than those for non-ELLs (Johnson et al., 2009; Vanderwood, Tung, & Checca, 2013). Results of one study suggested that students with lower language proficiency levels might need lower cut scores than non-ELLs, and students with higher language proficiency levels might need higher cut scores than non-ELLs (Vanderwood, Tung, & Checca, 2014). Another study suggested that cut scores for ELLs would need to be lower than those for non-ELLs (Johnson et al., 2009).

Finally, diagnostic accuracy results that were adjusted for base rate of pass/fail of the outcome measure were only reported in one study (Hosp, Hosp, & Dole, 2011). However, several studies reported lower base rates of passing of the included ELLs as compared to non-ELLs (Muyskens et al., 2009; Echols, 2011). Base rates are important to consider when examining diagnostic accuracy because outcomes that are more frequent, such as failure on a reading outcome measure, are more easily accurately predicted than outcomes that are less frequent, such as success on a reading outcome measure. If ELLs demonstrate lower passing rates on outcome measures, it may be more difficult to accurately predict outcomes for ELLs.

Limitations

The interpretation of the results of the present synthesis are limited to only those dissertations and peer-reviewed articles available. Because of the different terms used to describe students learning to read in more than one language, it is possible that the search terms were not comprehensive enough and some studies were not located. For example, a study that described students as Limited English Proficient (LEP) rather than ELL may not have been found. As with other systematic reviews, it is possible that the unpublished work not included yielded systematically different results. It is also possible that some publisher materials were not located. Also, the present review provides a systematic synthesis, not a quantitative synthesis or meta-analysis. Quantitative analyses of the results may provide more insight into the relationship between ORF and reading outcomes for emerging biliterates.

Future Directions

First, future research must not consider emerging biliterates as a homogeneous population. Biliterates, much like bilinguals, have different language proficiency and language backgrounds. These are influenced by language instruction type and content. Similar to ORF research conducted across grades, research must be conducted across different populations of emerging biliterates.

Next, future research must define and measure variables relevant to biliteracy. In order to build criterion-related or construct-related validity evidence, it is necessary to measure not just reading comprehension and ORF, but also vocabulary and oral language skills. Furthermore, these skills should be measured in both languages to better understand the compensatory process that is likely part of biliteracy development.

Finally, future research should consider alternatives to a single ORF as a screening tool. Evidence about the functioning of ORF may provide screening tools that could be added to ORF to improve accuracy. Multiple-gating screening procedures and local norms may allow ORF to be part of efficient, feasible, flexible screening for all students, regardless of language background.

Conclusion

There are numerous studies that provide validity evidence in support of the use of ORF as a screening tool for monolingual English speakers. The number of studies conducted with students learning to read in more than one language is far smaller. Results of the present synthesis suggest that ORF and reading outcomes may have a different relationship for emerging biliterates than for monolingual English speakers. Clearly, more research is needed to understand the functioning of ORF with emerging biliterates.

CHAPTER 3: Study Two

As awareness of the benefits of bilingualism increases (Bialystok, 2011), the number of dual language education programs seems to be increasing as well. In particular, there is an increasing number of immersion programs in the United States (T Fortune, personal communication, January 2, 2018). Regardless of language environment, early identification and intervention are key to the prevention of poor reading outcomes (Torgeson, 2002). Yet it may be difficult to differentiate between normative language acquisition and reading difficulties for immersion students (Durgunoglu, 2002; Genesee et al, 2013; Genesee & Fortune, 2014; Paradis, Genesee, & Crago, 2011). Therefore, it is important to have accurate reading screening assessments for the population of immersion students. Curriculum-based measurement of reading (CBM-R) oral reading fluency (ORF) is one possible relatively simple, low-inference assessment that may have the potential to distinguish between immersion students with normative language and literacy acquisition and those who may be at risk in reading.

As defined previously, bilinguals are individuals with sufficient language proficiency in more than one language, to "function in a situation that is defined by specific cognitive and linguistic demands, to a level of performance indicated by either objective criteria or normative standards" (Bialystok, 2001, p. 18). Bilinguals equally proficient in their first language (L1) and their second language (L2), or balanced bilinguals, are relatively rare (Bialystok, 1988, 2001; Grosjean, 2010).

Early theories of language acquisition espoused a subtractive-bilingualism view. That is, if a child learned a second language (L2) as well as their first language (L1), L2 would subtract from learning and development (de Valenzuela & Niccolai, 2004;

Bialystok, 2001). Many educational programs still aim to build proficiency in L2 without maintaining any proficiency in L1, particularly for language minority students, such as Spanish-speaking students learning English in the United States. These English-only programs in the U.S. are not dual language education programs, nor do they align with best practices in language education and acquisition (Christian, 2011; Francis, Lesaux, & August, 2006; Lindholm-Leary, 2008).

Research has now suggested that bilingualism can be a benefit (Bialystok 1988; Bialystok, 2011). For example, bilinguals demonstrate better cognitive or executive control than monolinguals, as measured by performance on the Stroop task (Bialystok, 2011). These cognitive benefits hold true from infancy through old age (Bialystok, 2011). Therefore, dual language education programs may benefit students.

Dual Language Education

Dual language education can be broadly defined as any type of education involving the regular use of two or more languages. More specifically, in the United States, these are languages where at least 50% of the day is spent learning in a language other than English. Dual language programs use two languages to provide instruction, and they aim to produce bilingual and biliterate students (Christian, 2011; Howard et al., 2007). Dual language education programs can be divided into four categories: (a) developmental bilingual programs, in which language minority students learn in their first language as well as their second, such as Spanish-speaking students in the US learning English; (b) two-way immersion programs, where half of the student population speaks the minority language and half speaks the majority language, such as Spanish/English programs with native Spanish and native English speakers, (c) heritage language immersion, such as students from indigenous cultures learning an endangered heritage language, and (d) foreign language immersion, in which students of a majority language are immersed in a foreign language, such as the native English speakers immersed in German in the present study (Christian, 2011; Howard et al., 2007).

Best practices in immersion instruction. Best practices in immersion instruction are informed by decades of research (Lyster & Tedick, 2014). The focus of immersion programs is to teach another language through teaching content in that language (Christian, 2011). For instance, students in an immersion setting learn grade-level content, like social studies, in a target language, like Spanish. Built in to each lesson are dual goals: acquisition of content and acquisition of target language.

Best practices in immersion instruction have evolved over the past several decades (Lyster & Tedick, 2014). Immersion instruction was originally thought to be most effective when students acquired language solely through comprehensible input or content in the target language (Howard et al., 2007; Krashen, 1981; Lyster & Tedick, 2014). Students would naturally acquire another language as a child might naturally acquire a first language, with little correction or explicit instruction. More recent reviews of evidence suggest that instruction should also include opportunities for comprehensible output, as well as explicit corrections and instruction in language to support formal language development (Christian, 2011; Lyster & Tedick, 2014). Therefore, bilingual programs should still provide at least 50% of the instruction in the second language (Christian, 2011), but should also set aside instructional time to discuss second-language grammar and cross-language similarities (Lyster & Tedick, 2014). Theories in literacy development also inform immersion instruction.

Literacy outcomes in immersion. Many of the earliest immersion programs, begun in the 1970s and 1980s, were for English-speaking students immersed in the French language in Canada (Genesee, 1987). Researchers sought to build evidence that these newly-created immersion language programs led to positive student outcomes (cf. Genesee et al., 1985). Specifically, researchers sought to verify that (a) immersion-educated students did not fall behind monolingual, English-speaking peers in literacy outcomes, and (b) immersion educated students achieved some measure of literacy in French.

Longitudinal studies of immersion students' English literacy suggest that upon beginning an early total one-way immersion program, English literacy lags behind the English literacy of peers in monolingual English programs (Swain & Lapkin, 1982). However, within one or two years of receiving English literacy instruction, or around fifth grade if English is introduced in third grade, immersion students' English literacy skills were comparable with English literacy skills of monolingual peers (Genesee et al., 1985; Genesee & Jared, 2008). More recent reviews of French immersion program outcomes suggest that at older grades, immersion students' English literacy may even exceed that of monolingual English peers (Genesee & Jared, 2008). A recent, comprehensive review of evidence on student outcomes in all types of immersion programs found a similar pattern: first-language literacy for immersion students is eventually comparable to or even exceeds the literacy of monolingual peers (Lindholm-Leary & Genesee, 2014).

Second-language literacy outcomes are related to the type of immersion program (Genesee, 1987). For example, students in full immersion, who are exposed to more of the target language, have better outcomes than students in partial immersion, and students who start immersion at a younger age have better second-language literacy outcomes than students who start immersion at an older age (Genesee, 1987; Howard et al., 2007). Immersion students do develop better literacy than monolingual peers in typical second-language instruction (Lindholm-Leary & Genesee, 2014).

Some of these positive outcomes for literacy in another language may be explained by a compensatory model of dual-language literacy acquisition. This model posits that students' literacy skills in the first language serve to compensate for emerging skills in the second language in an interactive manner (Bernhardt, 2005). In other words, some literacy skills transfer across languages in certain contexts. There is a body of evidence to suggest that some basic literacy skills such as phonological awareness do, in fact, transfer across languages (Durgunoglu, 2002; Bournot-Trites, 2008). For example, one seminal study with English-speaking students in French immersion investigated the cross-linguistic relationship between phonological awareness and decoding (Comeau et al., 1999). Results suggested that phonological skills in English or French were similarly related to decoding in English, and phonological skills in English or French were similarly related to decoding in French.

However, some skills may not transfer as readily. For example, in a longitudinal study of English-speaking students in a French early total immersion program, English vocabulary was a strong predictor of eventual English reading comprehension, but English vocabulary was not an important predictor of eventual French reading comprehension (Jared et al., 2011). Other studies have similarly found weak or no correlation between vocabulary and reading outcomes across languages (Lindsey, Manis, & Bailey, 2003; Verhoeven, 1994).

Individual differences in outcomes. There are always individual differences in educational outcomes. Educators have expressed concern that there are some students for whom immersion education would be harmful, such as students with low cognitive abilities or students with learning disabilities (Genesee & Fortune, 2014). However, recent summaries of immersion research (Genesee et al., 2013; Genesee & Fortune, 2014) suggest that cognitive ability is correlated to language proficiency primarily for older students. Results of one study (Myers, 2009, as cited in Genesee & Fortune, 2014) suggested that students with special needs in a two-way immersion school had similar performance to students with special needs in a monolingual school across literacy, math, and science. Despite these initial results in support of immersion for diverse learners, there are relatively few assessment tools that can accurately identify immersion students who are at-risk in literacy (Fortune, with Menke, 2010; Genesee et al., 2013; Genesee & Fortune, 2014).

Universal Screening of Reading

Universal screening of reading involves the assessment of all students on selected reading skills (Jenkins, Hudson, & Johnson, 2007). Then, targeted interventions can be provided to students whose assessment results suggest they are at-risk in reading. Evidence suggests that early identification and intervention are important to prevent poor reading outcomes (Torgeson, 2002).

For monolingual English students, a large body of evidence supports the use of curriculum-based measurement of reading (CBM-R) oral reading fluency (ORF) as a screener. ORF is simply administered; a student reads aloud for one minute, and an examiner counts the number of words read correctly by the student (Deno, 1985). The resulting score of words read correctly per minute (WCPM) has been shown to be both a good indicator of overall reading (Deno, 2003; Fuchs, 2004; Reschly et al., 2009; Wayman et al., 2007), as well as an accurate predictor of students' future reading outcomes on both reading achievement assessments as well as state accountability assessments (Kilgus et al., 2014). Multiple-gating procedures, or using a series of tools to screen reading skills, may improve prediction accuracy further (Compton, Fuchs, Fuchs, & Bryant, 2006; Johnson et al., 2009; Glover & Albers, 2007).

There remain relatively few studies that provide evidence in support of ORF with bilingual students learning to read, or emerging biliterates. Those available studies demonstrate inconclusive results. Several studies suggest that for emerging biliterates, ORF tasks correlate with reading outcome tasks with similar magnitude as for monolingual English speakers (Baker & Good, 1995; Wiley & Deno, 2005). However, results of several other studies suggest that some English language learners may be able to read fluently, but that ORF measures may not accurately predict reading comprehension (Knight-Teague, Vanderwood, & Knight, 2014; Quirk & Beem, 2012). This suggests that ORF may not function as a measure of reading comprehension for emerging biliterates, but rather may function as solely a measure of decoding or word-level reading skill.

Reading screening in immersion environments. Screening for reading difficulties in the immersion environment is complex for several reasons. First, students acquiring a language may appear similar to students with reading difficulties (Chu &

Flores, 2011; Durgonuglu, 2002; Genesee et al., 2013; Genesee & Fortune, 2014; Paradis, Genesee, & Crago, 2011). For example, a student with limited language proficiency may have poor listening comprehension, difficulty following directions, and may make frequent language errors. Students with learning disabilities may display many of these similar traits (Durgonuglu, 2002; Chu & Flores, 2011). It is inappropriate, however, to wait for advanced language proficiency to assess for learning disabilities (Paradis, Genesee, & Crago, 2011). As with monolingual students, early identification could lead to early intervention and better outcomes. Therefore, it is important to understand the learning trajectory of students in the immersion setting, just as in any other instructional setting.

Second, reading screening in immersion environments is difficult because there is a distinct lack of evidence for many reading screening tools commonly used for monolingual populations (Genesee, 2007; Genesee et al., 2013; Genesee & Fortune, 2014). Some studies have begun to build evidence for the cross linguistic transfer of phonemic awareness skills and, correspondingly, the cross-linguistic use of phonemic awareness screening assessments. For example, one study investigated early reading screening tools for English-speaking French immersion students (MacCoubrey, Wade-Woolley, Klinger, & Kirby, 2004). Results suggested that phoneme blending in English (L1) could differentiate between poor and typical readers in English as well as French. Another study found that English phonological awareness in Kindergarten was a predictor of French reading comprehension in first grade (Erdos et al., 2010). However, cross-language reading comprehension or broad reading outcome screening evidence is more limited.

There is some evidence that use of first-language assessments may be the most appropriate for students with minimal second language proficiency (Genesee et al., 2013). ORF, a simple, low inference reading assessment, may function well within the immersion environment (Fortune, 2011). However, to date, there are no known studies of the use of ORF as an indicator of reading in immersion schools, nor are there any known studies of the use of ORF to identify immersion students who are at risk of not meeting reading standards (Genesee et al., 2013).

Purpose

The purpose of this study was twofold. First, the study sought to evaluate the extent to which oral reading fluency measures in English and German are predictive of reading outcomes for German immersion emerging biliterates. Specifically, to what extent is measurement of ORF in English and German predictive of English reading success, as measured by Minnesota Comprehensive Assessment Reading performance? Do the predictive properties of German and English ORF change across grades three to five? To what extent does German language proficiency, as measured by the oral communication, listening comprehension and reading comprehension portions of a German language proficiency assessment (IVA A1), mediate the relationship between English ORF and English reading success? To what extent does German language proficiency, as measured by the oral communication and listening comprehension portions of a German language proficiency assessment (IVA A1) mediate the relationship between German ORF and German reading success? To what extent does English (L1) ORF describe variability in English reading over and above what German (L2) ORF can describe? To what extent does German (L2) ORF describe variability in German reading comprehension over and above what English (L1) ORF can describe?

Second, this study sought to evaluate the ability of German and/or English ORF to accurately identify students who are at risk in reading. Specifically, how accurately can either German ORF, English ORF, or a combination of the two predict which students will not meet proficiency on the MCA?

Methods

Participants

Participants were all native English speakers learning German in one immersion school. The school was recruited through a community partnership with the researcher. In exchange for support with implementation of screening measures and work toward adopting multi-tiered system of service delivery, extant data was shared by the school. Therefore, inclusion in the study was based on passive consent. Parents were informed of the new school screening policy and were given the choice to opt out of the study's extra data analysis.

The student population (*n* = 520) at the school is primarily White; 88% of students identify as White, 2% identify as Black, 2% identify as Asian, 1% as Hispanic, and 8% as two or more races. Only 5% of students participate in free or reduced-price lunch programs, 8.8% of students receive special education services of some type, and 2% of students are English Language Learners.

Students whose first language was not English were excluded from the final analysis sample. The final sample included data from students in third (*n* = 60), fourth (*n*

= 60), and fifth grade (*n* = 42). The school begins explicit literacy instruction in English in third grade, and therefore begins using English screening measures in third grade.

Setting

The immersion school is a public charter school, and parents undergo a random lottery process to enroll students. Students are strongly discouraged from enrolling after Kindergarten unless they have been attending a German-language school. The instructional model strives for 90% of instruction in German from Kindergarten to third grade. In third grade, literacy instruction in English begins. By fifth grade, some special subjects such as science and math are taught in English as well as German. In sixth through eighth grades, the program has approximately 50% of instruction in German; math and language arts are taught in English, science is taught in English and German, and language arts, social studies, and specials (physical education, art) are taught in German.

At the time of this study, the school implemented English ORF screening in grades three, four, and five with measures from FastBridge Learning (FBL; TJCC, 2015). During the 2016-2017 school year, third grade students were screened three times: fall, winter, and spring. Fourth and fifth grade students were screened winter and spring. German screening measures were used three times per year from kindergarten through fifth grade. Screening measures in both languages were intended to be used to identify students in need of extra services.

Measures

Four measures were used in analyses to address research questions: (a) ORF in German, (b) ORF in English, (c) the Internationale Vergleichsarbeit (IVA; ZFA, 2014), and finally (d) the Minnesota Comprehensive Assessment in Reading (MCA; MDE 2015). These assessments are described in the following paragraphs.

Oral reading fluency in English. Given that the school already employs ORF using FBL, the current study also used this assessment tool. ORF within FBL consists of three oral reading fluency (ORF) probes per screening session. For each probe, a student reads a grade-level passage for one minute. Words read correct are counted. The final reported score is the median of the three probes. FBL reports classification accuracy of above .80 in grades 3, 4, and 5 when comparing CBM-ORF performance to performance on a reading assessment (the Test of Silent Reading Efficiency and Comprehension; TOSREC). Median test-retest and alternate form reliability is reported to be above .90, as is inter-rater reliability. Concurrent validity coefficients of FBL ORF and the TOSREC are .81, .79, and .81 for third, fourth, and fifth grades respectively. Predictive validity coefficients of FBL ORF and the TOSREC are .69, .52, and .87 for third fourth, and fifth grades respectively.

Oral reading fluency in German. One probe was administered to each student during each screening period. The examiner measured words read correct per minute as with the English CBM ORF measure. To measure reading loss over the summer, the German probes remain consistent from spring to fall screening periods. For example, a second-grade student in the spring would read a text, and then return in the fall as a third grader and read that same text.

German reading fluency probes were developed by a native German speaker with a doctorate in German language. Texts were pulled from curricular materials at the immersion grade level and adapted as needed. For example, the 3rd grade reading passage was pulled from materials used in the 3rd grade classroom at the school. Students at this immersion school typically read passages one or two grade levels below what monolingual German speaking students might read. Therefore, 3rd grade students at the immersion school read texts categorized to be at the 1st or 2nd grade level by monolingual German educators. All passages selected were narrative passages about experiences that were judged to be familiar to the English-speaking students in the school. For example, one passage described the receipt of a birthday present, and another passage described visiting the garden of a family member.

Although students may demonstrate different levels of performance on curriculum texts than on commercially-developed texts, technical adequacy of the level score is still likely to be acceptable (Wayman et al., 2007). The German language is more orthographically shallow than English and has fewer allowable syllable sounds, but it has a similar number of syllables per word (Pellegrino, Coupe, & Marsico, 2011). However, German grammar and sentence construction can be more complex, resulting in longer sentences. Because previous research suggests that readability indices are not good indicators of passage difficulty (Ardoin et al., 2005), passages were selected with a focus on simple sentences and shorter words. The English measure used in the present study (FBL ORF) was developed based on word length and sentence length rather than readability indices (TJCC, 2015). An analysis of text complexity and readability of the three German passages is presented in Table 3. One of the most commonly used measures of readability in the German language is the Lesbarkeitsindex, or LIX (Björnsson, 1968, as cited in Klare, 1984). The LIX is calculated by adding the average number of words per sentence to the percent of long words (words above 6 letters). The number of syllables in each passage and the resulting average number of syllables per word was also calculated. Results in Table 3 suggest that the 5th grade passage was more difficult than the 4th grade passage, and the 4th grade passage was more difficult than the 3rd grade passage.

Table 3.

Readability and other passage difficulty variables for each grade-level German reading passage.

Grade	Word count	Syllable Count	Sentence Count	Sentence Length	Syllables per word	Percent Long Word	LIX	Descriptor
3	100	161	15	6.6	1.61	10	16.6	Very Easy – First Grade
4	236	366	40	5.9	1.55	17.3	23.2	Very Easy – First Grade
5	426	641	52	8.1	1.50	16.6	24.8	Easy – Second Grade

LIX = Lesbarkeitsindex (Readability Index)

The Internationale Vergleichsarbeit (IVA). The IVA (International Comparative Work) is one assessment under the broad umbrella of the Deutsches Sprachdiplom (DSD; German Language Diploma). The DSD is a diploma system created by German education and foreign office initiatives to promote the German language. The assessments within the diploma system are given only in schools registered with the German government, where non-native German speakers are learning German. The final DSD exam, the DSDII, results in a German language proficiency rating based on the Common European Framework of Reference of Languages, a common way to rate language proficiency in European nations. This rating can allow German speakers from outside of Germany to demonstrate sufficient language proficiency to attend German University classes.

Two exams allow international schools to determine whether their German-learning students are making appropriate progress toward the DSDII: IVA A1 and IVA A2. At the German immersion school included in the study, IVA A1 is given to students in third grade, to determine whether they are on track to be measured proficient by the time they take the DSD. Administrators of the IVA must be proficient in German and undergo a brief training.

The IVA measures listening comprehension, reading comprehension, writing, and speaking in German. The tasks germane to the present study are speaking (Mündliche Kommunikation; MK), listening comprehension (Hörverstehen; HV) and reading comprehension (Leseverstehen; LV).

The speaking or oral communication task takes approximately 15 minutes. During the first portion, students are asked to speak as a monologue on a particular every day topic and are given cue words they must incorporate into their speaking. Students are scored based on their skill in speaking with connected, fluent language. In the second portion, students are asked to engage in a dialogue with a fellow student, and they are scored on their capacity to respond to questions with connected, fluent language.

The reading comprehension task takes approximately 15 minutes. Students are asked to read short (100 word) narrative texts, and answer five true/false questions. Next, students are asked to complete five sentences as a cloze exercise.

The listening comprehension task also takes approximately 15 minutes. Students are asked to listen to a brief (100 word) narrative text, read aloud by the examiner, and are asked to complete five sentences about the text. Finally, students are asked to demonstrate comprehension nonverbally by drawing key elements from a narrative text.

Raw scores are reported, as well as a proficiency level. For example, a student taking the IVA A1 exam who met a cut score would be reported to be proficient at level A1. A student who did not meet the cut score would be reported to be not proficient at level A1.

Based on comparisons of raw scores and proficiency designations, the cut scores seem to change across years. Information on the norming and development, as well as psychometric properties of the assessment, were requested from a German government representative from the Zentralstelle für das Auslandsschulwesen (Central Office for Foreign Schools); no further information was available. Although the tasks required as part of the IVA A1 exam align with tasks included in other language proficiency exams, there is insufficient evidence to suggest that the IVA A1 language proficiency exams are well validated language proficiency measures. Therefore, results that include IVA A1 scores should be interpreted with caution.

Minnesota Comprehensive Assessment (MCA). The MCA is a state accountability test developed by the Minnesota Department of Education in conjunction with measurement and content experts (MDE, 2015). The MCA is given to students in grades three through eight, as well as students in grade ten, and is designed to measure proficiency in Minnesota standards in math, reading, and science. The most recent iteration of the MCA of Reading, the MCA Reading Series III, was created in 2013 to align with the 2010 Minnesota K-12 Academic Standards in English Language Arts (ELA). The assessment is given primarily online, with several examiners monitoring a class of students. Items consist of literature and informational text reading, with corresponding questions designed to measure analysis, interpretation, and evaluation of text. Separate passages and questions are given at each grade level, with the exception of a few linking items to aid in vertical alignment of scores. Scores are calculated within an Item Response Theory (IRT) framework, and then transformed to easily interpretable scaled scores. Scaled scores range from G01 to G99, where G represents the student's grade. For example, a fourth-grade student could receive a score of 449, whereas a fifth grade student could receive a score of 549. Because of careful linking and equating procedures, some comparison of scores across grades is possible. Specifically, a score of G50 is the cutoff for the designation of "Meets Standards," and G40 is the cutoff for the designation of "Partially Meets Standards" across grades.

The technical manual presents evidence to support the MCA Reading's proposed interpretation and use as a measure of student proficiency on Minnesota state ELA standards (MDE, 2015). Construct-related validity evidence is strong, based on thorough test specification, item creation, and field testing procedures. Criterion-related validity evidence is missing, as no studies have been conducted comparing the MCA Reading to other assessments of reading. Evidence of consistency, as measured by internal consistency, is strong (.91; MDE, 2017) across grades three, four, and five.

Procedure

All students in grades three, four, and five received German ORF three times per year. Students in grade three received English ORF three times per year. Students in grades four and five received English ORF only during winter and spring. All third-grade students took the IVA A1 at the end of the year. Students across all grades took the state accountability assessment (MCA) at the end of the school year.

Analysis

Within each grade, the following procedure was conducted to prepare data and understand the included variables. First, descriptive information was gathered about the demographic variables of the participants, including special education eligibility and language background. As stated previously, non-native-English-speaking students were excluded from the study. Next, the assessment variables were explored, including descriptive information, as well as correlations among ORF English, ORF German, IVA A1 oral communication, IVA A1 reading comprehension, IVA A1 listening comprehension, and MCA Reading.

Based on the results of the data exploration described, three more analyses were conducted at each grade level. First, a single-level regression model was constructed in a step-wise fashion to understand the relationship between predictor variables of German proficiency, ORF in both languages, and an English reading outcome measure. Next, single-level regression models were conducted to investigate the extent to which English ORF and German ORF explained cross-language reading proficiency (Baker, Park, & Baker, 2012). Finally, a Receiver Operating Curve analysis was conducted to determine the accuracy of predictions. These analyses are described in detail below.

Single-level regression models. A single level-regression model at each grade allowed exploration of which variables best explain variability in the outcome measure. In other words, to what extent do reading screening measures explain variability in the MCA, accounting for potential differences based on demographics and German language proficiency? The following variables were added to the linear model to predict outcome on the MCA reading end-of year assessment: (a) student special education status, coded as 0, not receiving special education services, and 1, receiving special education services, based on data in the school's student information system; (b) German oral communication as measured by the IVA, a continuous variable from 0 to 10; (c) German listening comprehension as measured by IVA, a continuous variable from 0 to 10; (d) German reading comprehension as measured by the IVA, a continuous variable ranging from 0 to 10 (e) ORF German WRCPM in winter, (f) ORF English WRCPM in winter. Although a research question suggested a similar single level regression model with German reading comprehension as the outcome variable, there was insufficient variability in the German reading comprehension measure for it to be used as a dependent variable in this type of multi-predictor regression.

Cross-language explanation of reading outcomes. At each grade level, two series of regression analyses were conducted. First, German reading comprehension as measured by the IVA A1 was predicted by English ORF, German ORF, and then both English ORF and German ORF. Then, English reading skill as measured by the MCA reading end-of-year assessment was predicted by German ORF, English ORF, and then both German ORF and English ORF. These analyses provided information about the extent to which ORF probes can predict reading success across languages. For example, does English ORF still predict German reading comprehension after German ORF is added to the model, or does German ORF then act as the better predictor of German reading comprehension?

Diagnostic accuracy analysis. Diagnostic accuracy analyses were conducted via Receiver Operating Curve (ROC) analysis. The results of the single-level regression models at each grade level determined the predictors in the diagnostic accuracy analysis at each grade level. For example, should English ORF explain the most variability in fourth grade MCA results, a logistic regression would be built with English ORF as the predictor and MCA proficiency (G50 or "Meets Standards") as the outcome. The predicted probabilities of that logistic regression allowed a ROC plot, which can help determine ideal cutoff scores and maximize sensitivity and specificity.

Power analysis. Before data collection, a power analysis was conducted. The participating school had approximately 70 students per grade. Previous literature (cf. Wiley & Deno, 2005) suggests that an effect size of .2 is a conservative estimate for variability in MCA scores explained by the eight initially proposed predictors. Power was therefore estimated to be approximately .82 based on the proposed six predictors (Champely, 2016).

After data collection and data cleaning, another power analysis was conducted based on the actual sample sizes and six included predictors. Based on a sample of 60 students, and an effect size of .2, power was estimated to be .86. Based on the smaller sample of 42 fifth grade students, power was estimated to be .69.

Results

First, results are presented about the relationship between ORF and reading outcome measures in each grade. Then, the results of diagnostic accuracy analyses are presented for each grade.

Relationship Between Reading Fluency Measures and Outcome Measures

The first purpose of the present study was to evaluate the extent to which oral reading fluency measures in English and German are predictive of reading outcomes in English and German for emerging biliterates in the German immersion setting. Descriptive statistics, correlational analyses, and regression analyses were conducted to answer research questions. Assumptions of linearity, independence, normality, and homoscedasticity were examined via scatterplots, Durbin Watson tests, investigation of descriptive statistics, and plots of model residuals. Unless otherwise stated below, all models met assumptions necessary for analyses.

Grade 3. The variables used in the third-grade analysis are described in Table 4. These variables include English and German ORF scores, MCA Reading scores, as well as German language proficiency assessment including reading, listening, speaking and writing. The mean performance on English ORF was 111.95 words read correct per minute (*SD* = 45.89). German reading fluency was measured to be lower, with a mean of 68.4 (*SD* = 28.24). Average MCA Reading performance was close to the middle of the scale, at 349.15 (*SD* = 21.74). Mean scores on variables measuring German language proficiency were fairly high, with listening comprehension the highest (*M* = 9.83, *SD* = 0.83). Most variables included in regression analysis are approximately normally distributed, with the exception of the German listening comprehension measure (A1HV), which is negatively skewed.

Table 4.

Descriptive statistics of relevant third grade variables.

	n	average	SD	MDN	MIN	MAX	skew
English WRC	60	111.95	45.89	116.5	16	182	-0.44
English Accuracy	60	96.8	5.23	98.64	72.73	100	-2.51
A1LV	60	7.53	1.87	8	3	10	-0.25
A1HV	60	9.83	0.83	10	4	10	-6.06
A1SK	60	7.52	1.42	7	3	10	-0.38
A1MK	60	7.55	1.5	8	0	10	-2.15
German WRC	60	68.4	28.24	76.5	5	138	-0.27
German Accuracy	60	0.93	0.11	0.97	0.29	1	-3.62
MCA Reading	60	349.15	21.74	352	301	390	-0.44

WRC – Word read correct, A1LV – German reading comprehension, A1HV – German listening comprehension, A1SK – German written communication, A1MK – German oral communication

Correlations among third grade variables are displayed on Table 5. There is a strong, significant correlation between MCA reading scores and English ORF scores ($r = 0.84$, $p < 0.00$). There is also a strong, significant correlation between MCA reading scores and German ORF scores ($r = 0.70$, $p < 0.00$). Other correlations of interest include a moderate, significant correlation between German reading comprehension and MCA performance ($r = 0.63$, $p < 0.00$). Oral reading fluency across languages also is highly correlated ($r = 0.89$, $p < 0.00$).

Table 5.

Correlations among third grade variables of interest. N=60

	1	2	3	4	5	6	7	8	9
1 EWRC	1.00								
2 E. Acc	0.74***	1.00							
3 A1LV	0.53***	0.34**	1.00						
4 A1HV	0.26*	0.26*	0.15	1.00					
5 A1SK	0.46***	0.32*	0.34**	0.20	1.00				
6 A1MK	0.32*	0.19	0.29*	0.03	0.24	1.00			
7 GWRC	0.89***	0.68***	0.41**	0.11	0.46***	0.33**	1.00		
8 G Acc	0.57***	0.78***	0.32*	-0.04	0.27*	0.28*	0.65***	1.00	
9 MCAR	0.84***	0.65***	0.63***	0.25	0.42***	0.27*	0.70***	0.47***	1.00

* $p < .05$ ** $p < .01$ *** $p < .001$

EWRC – English WRC, E. Acc – English Accuracy, A1LV – German reading comprehension, A1HV – German listening comprehension, A1SK – German written communication, A1MK – German oral communication, GWRC – German WRC, G Acc – German Accuracy, MCAR – Minnesota Comprehensive Assessment – Reading.

The initial review of correlations suggests that English and German ORF probes may be good predictors of performance on the MCA reading assessment. Results of the step-wise linear regression models (Table 6) suggest that students in special education are likely to obtain lower scores on the MCA reading assessment. However, the Special Education variable becomes non-significant once reading fluency scores are added to the model. In other words, these results suggest that the lower performance by students receiving Special Education services is well-captured by reading fluency probes.

German reading comprehension and German ORF are both significant predictors of MCA reading scores. Once English ORF is added in to the final model however, German ORF is no longer a significant predictor of MCA scores. However, German reading comprehension remains a small yet significant predictor. Based on the high amount of variability explained by the final model (R^2 = 0.76), these results suggest that English ORF is a strong predictor of MCA reading scores in the third-grade immersion setting.

Table 6.

Third grade step-wise regression model building, predicting end of year MCA reading assessment.

	Grade 3 Model 1	Grade 3 Model 2	Grade 3 Model 3	Grade 3 Model 4	Grade 3 Model 5	Grade 3 Model 6
Intercept	353.27 (2.87)	324.54 (34.25)	300.42 (37.62)	279.68 (31.28)	258.41 (26.11)	287.62 (22.55)
Special Education	-22.45** (6.70)	-20.15** (7.25)	-17.19* (7.45)	-13.27* (6.19)	-1.02 (5.62)	2.64 (4.74)
A1HV	-	2.88 (3.42)	3.24 (3.39)	1.96 (2.81)	3.24 (2.33)	0.40 (2.02)
A1MK	-	-	2.65 (1.80)	0.63 (1.53)	-0.46 (1.28)	-0.21 (1.07)
A1LV	-	-	-	6.35*** (1.22)	4.64*** (1.06)	2.67** (0.96)
GWRCW	-	-	-	-	0.40*** (0.08)	-0.13 (0.13)
EWRCW	-	-	-	-	-	0.42*** (0.09)
R^2	0.16	0.17	0.20	0.47	0.65	0.76

* *p* < .05 ** *p* < .01 *** *p* < .001

EWRC – English WRC, A1LV – German reading comprehension, A1HV – German listening comprehension, A1SK – German written communication, A1MK – German oral communication, GWRC – German WRC.

Table 7 displays the results of the cross-language analyses predicting to the English outcome measure (MCA Reading). As also suggested by the correlation and regression analyses already described, English ORF in isolation is a significant predictor of MCA Reading scores (R^2 = 0.71). German ORF in isolation is also a significant predictor of MCA Reading scores, yet it describes less overall variability than the English ORF scores (R^2 = 0.49). When combined in a single regression model, only the English ORF scores were significant predictors of MCA Reading scores.

Table 7.

Cross-language regression results with English MCA as outcome variable.

	Grade 3	Grade 3	Grade 3	Grade 4	Grade 4	Grade 4	Grade 5	Grade 5	Grade 5
	Model 0	Model 1	Model 2	Model 0	Model 1	Model 2	Model 0	Model 1	Model 2
Int.	304.59	312.37	305.95	417.33	428.71	417.39	534.91	547.28	536.39
	(4.08)	(5.35)	(4.10)	(5.17)	(4.88)	(5.20)	(8.00)	(8.22)	(8.33)
G	-	0.54***	-0.20	-	0.36***	0.04	-	0.18*	-0.07
WRC		(0.07)	(0.12)		(0.06)	(0.10)		(0.08)	(0.11)
E	0.40***	-	0.51***	0.30***	-	0.27***	0.19***	-	0.22**
WRC	(0.03)		(0.07)	(0.04)		(0.07)	(0.05)		(0.07)
R^2	0.71	0.49	0.72	0.53	0.41	0.53	0.27	0.10	0.27

* $p < .05$ ** $p < .01$ *** $p < .001$

EWRC – English WRC, GWRC – German WRC.

Table 8 displays the results of the cross-language analyses predicting to the German outcome measure (A1LV). Again, as suggested by correlations, German ORF scores were significant predictors of German reading comprehension in isolation, yet they explained a relatively small amount of variability (R^2 = 0.17). In isolation, English ORF scores were slightly stronger predictors of German reading comprehension (R^2 = 0.27). When English and German ORF were combined in a model, neither predictor was significant, and a relatively small amount of variability was explained (R^2 = 0.31). The smaller amount of variability explained may also be related to the smaller amount of variability in the German reading comprehension variable. Rather than a range of G00 to G99 on the MCA Reading score, the German reading comprehension variable ranged only from 0 to 10 and was negatively skewed as well. Furthermore, as previously described, there is limited information to support the validity of the German language proficiency assessment.

Table 8.

Cross-language regression results with German reading comprehension (A1LV) as outcome measure.

	Grade 3 Model 0	Grade 3 Model 1	Grade 3 Model 2	Grade 4 Model 0	Grade 4 Model 1	Grade 4 Model 2	Grade 5 Model 0	Grade 5 Model 1	Grade 5 Model 2
Int.	5.67	5.10	5.25	5.07	4.36	4.40	7.06	6.71	6.57
	(0.59)	(0.44)	(0.55)	(0.87)	(1.04)	(1.04)	(1.57)	(1.67)	(1.75)
E	-	0.02***	0.03	-	0.03**	0.02	-	0.01	0.01
ORF		(0.02)	(0.01)		(0.01)	(0.01)		(0.01)	(0.02)
G	0.03**	-	-0.02	0.04***	-	0.02	0.02	-	0.01
ORF	(0.01)		(0.02)	(0.01)		(0.02)	(0.02)		(0.02)
R^2	0.17	0.27	0.31	0.21	0.21	0.23	0.03	0.04	0.04

* *p* < .05   ** *p* < .01   *** *p* < .001

EWRC – English WRC, GWRC – German WRC.

Grade 4. The variables used in the fourth-grade analysis are described on Table 9. These variables include English reading fluency, German reading fluency, German proficiency measures, and MCA Reading scores. The average performance on the English ORF task was 137.75 words read correctly per minute (*SD* = 33.65), whereas the German was again lower, with an average of 81.68 words read correctly per minute (*SD* = 24.14). As with third grade, the average MCA Reading score was close to the middle of the score range (*M* = 458.15, *SD* = 13.67). Mean scores on the German proficiency measures were fairly high. Listening comprehension was again the highest (*M* = 9.52, *SD* = 1.98). As with the third-grade variables, many were normally distributed. The German listening comprehension (A1HV) was again negatively skewed, as students tend to do very well on this task.

Table 9.

Descriptive statistics of relevant fourth grade variables.

	n	average	SD	MDN	MIN	MAX	skew
English WRC	60	137.75	33.65	142.5	19	195	-0.80
English Accuracy	60	99.13	2.38	100	82.61	100	-5.66
A1LV	60	8.33	2.11	9	2	10	-1.37
A1HV	60	9.52	1.32	10	2	10	-4.29
A1SK	60	7.50	1.98	8	0	10	-1.55
A1MK	60	8.10	1.76	8	0	10	-1.83
German WRC	60	81.68	24.14	85	9	130	-0.54
German Accuracy	60	0.96	0.06	0.98	0.60	1	-3.70
MCA Reading	60	458.15	13.67	459.5	411	481	-0.86

A1LV – German reading comprehension, A1HV – German listening comprehension, A1SK – German written communication, A1MK – German oral communication

Correlations among variables are displayed on Table 10. As in third grade, there is a strong, significant correlation between English ORF and MCA Reading scores (r = 0.73, p < 0.00). There is a slightly smaller yet still significant correlation between German ORF and MCA Reading scores (r = 0.64, p < 0.00). English and German ORF scores again correlate (r = 0.84, p < 0.00). Other correlations of interest include moderate correlations of German listening comprehension and MCA Reading (r = 0.58, p < 0.00), which is larger than the correlation in the third grade, and German reading comprehension and MCA Reading (r = 0.56, p < 0.00), which is similar to the correlation in the third grade.

Again, correlations suggest that German and English ORF scores may be good predictors of MCA Reading results. The fourth-grade regression analysis results are presented on Table 11. Compared to the third-grade final model, the final fourth-grade model explained less variability in MCA Reading scores (R^2 = 0.63). Similar to the third-grade modeling, German ORF was a significant predictor until English ORF was added.

Also similar to the third-grade final model, German reading comprehension remained a significant predictor in the final fourth-grade model. As correlations would suggest, German listening comprehension was a significant predictor of MCA Reading scores in the final fourth-grade model. The amount of variability in MCA Reading scores explained by the final model is moderate to large ($R^2 = 0.63$), again suggesting that English ORF scores are a good predictor of MCA Reading scores in the fourth-grade immersion setting.

Table 10.

Correlations among fourth grade variables of interest. N=60

	1	2	3	4	5	6	7	8	9
1 EWRC	1.00								
2 E. Acc	0.60***	1.00							
3 A1LV	0.46***	0.42***	1.00						
4 A1HV	0.52***	0.72***	0.46***	1.00					
5 A1SK	0.40**	0.37**	0.43***	0.52***	1.00				
6 A1MK	0.22	0.23	0.45***	0.38**	0.74***	1.00			
7 GWRC	0.84***	0.49***	0.46***	0.51***	0.46***	0.32*	1.00		
8 G Acc	0.65***	0.75***	0.40**	0.65***	0.50***	0.32*	0.68***	1.00	
9 MCAR	0.73***	0.63***	0.56***	0.58***	0.41**	0.24	0.64***	0.58***	1.00

* $p < .05$ ** $p < .01$ *** $p < .001$

EWRC – English WRC, E. Acc – English Accuracy, A1LV – German reading comprehension, A1HV – German listening comprehension, A1SK – German written communication, A1MK – German oral communication, GWRC – German WRC, G Acc – German Accuracy, MCAR – Minnesota Comprehensive Assessment – Reading.

Table 11.

Fourth grade step-wise regression model building, predicting end of year MCA English reading assessment.

	Grade 4 Model 1	**Grade 4 Model 2**	**Grade 4 Model 3**	**Grade 4 Model 4**	**Grade 4 Model 5**	**Grade 4 Model 6**
Intercept	459.65 (1.83)	402.11 (12.30)	401.37 (13.57)	396.20 (12.46)	399.76 (11.37)	395.58 (10.60)
Special Education	-11.30* (5.02)	-0.95 (4.82)	-0.70 (5.21)	2.58 (4.85)	2.16 (4.41)	1.41 (4.08)
A1HV	-	5.90*** (1.25)	5.86*** (1.29)	4.66*** (1.23)	3.19** (1.19)	2.47* (1.13)
A1MK	-	-	0.13 (0.98)	-0.68 (0.92)	-0.92 (0.84)	-0.44 (0.79)
A1LV	-	-	-	2.73*** (0.78)	2.06** (0.73)	1.69* (0.69)
GWRCW	-	-	-	-	0.22*** (0.06)	-0.00 (0.09)
EWRCW	-	-	-	-	-	0.21** (0.07)
R^2	0.08	0.34	0.34	0.46	0.56	0.63

* $p < .05$ ** $p < .01$ *** $p < .001$

EWRC – English WRC, E. Acc – English Accuracy, A1LV – German reading comprehension, A1HV – German listening comprehension, A1SK – German written communication, A1MK – German oral communication, GWRC – German WRC, G Acc – German Accuracy, MCAR – Minnesota Comprehensive Assessment – Reading.

Table 7 displays the results of the cross-language analyses predicting to the English outcome measure (MCA Reading). Similar to third grade results, English ORF in isolation is a significant predictor of MCA Reading scores, yet less variance is explained by English ORF in fourth grade (R^2 = 0.53). German ORF in isolation is also a significant predictor of MCA Reading scores, yet similar to third grade, it describes less overall variability than the English ORF scores (R^2 = 0.41). When combined in a single regression model, only the English ORF scores were significant predictors of MCA Reading scores. German ORF did not add any explanatory power to the model.

Table 8 displays the results of the cross-language analyses predicting to the German outcome measure (A1LV). Again, as suggested by correlations, German ORF scores were significant predictors of German reading comprehension in isolation, yet they explained only a moderate to small amount of variability ($R^2 = 0.21$). In isolation, English ORF scores were a very similar predictor of German reading comprehension ($R^2 = 0.21$). When English and German ORF were combined in a model, neither predictor was significant, and a small amount of variability was explained ($R^2 = 0.23$). As in third grade, the smaller amount of variability explained of the German reading comprehension measure may be related to the smaller absolute variability in the variable, or psychometric properties of the assessment.

Grade 5. The variables used in the fifth-grade analyses are presented on Table 12. Again, variables included English reading fluency, German reading fluency, measures of German language proficiency, as well as MCA Reading scores. The average English ORF score was higher than average scores in third or fourth grade (*M* = 160.48*, SD* = 28.30). The German ORF score (M = 98.05*, SD* = 18.51*)* was lower than the English, similar to the pattern in third and fourth grades. Similar to variables described in third and fourth grades, German listening comprehension was the highest score of the German proficiency measures (*M* = 9.93, *SD* = 0.26) and was negatively skewed. It is important to note that the sample size in the fifth-grade analyses is smaller than the sample size of fourth- or third-grade analyses. This sample size difference and resultant lack of sufficient power may explain some of the smaller correlations (Table 13).

Table 12.

Descriptive statistics of relevant fifth grade variables.

	n	average	SD	MDN	MIN	MAX	skew
English WRC	42	160.48	28.30	162.00	94.00	216	-0.15
English Accuracy	42	99.27	0.87	99.46	97.04	100	-0.99
A1LV	42	8.86	1.87	10	3	10	-1.57
A1HV	42	9.93	0.26	10	9	10	-3.21
A1SK	42	8.55	1.02	8.50	7	10	0.01
A1MK	42	8.98	0.81	9	7	10	-0.49
German WRC	42	98.05	18.51	101	54	135	-0.54
German Accuracy	42	0.98	0.03	0.99	0.88	1.00	-1.72
MCA Reading	42	564.64	10.19	566	534	582	-0.64

A1LV – German reading comprehension, A1HV – German listening comprehension, A1SK – German written communication, A1MK – German oral communication

Table 13.

Correlations between fifth grade variables of interest. N=42

	1	2	3	4	5	6	7	8	9
1 EWRC	1.00								
2 E. Acc	0.20	1.00							
3 A1LV	0.20	0.11	1.00						
4 A1HV	0.40**	0.00	-0.17	1.00					
5 A1SK	0.33*	0.08	0.27	0.15	1.00				
6 A1MK	0.05	0.11	0.42**	0.11	0.31*	1.00			
7 GWRC	0.74***	0.29	0.18	0.51***	0.25	0.09	1.00		
8 G Acc	0.42**	0.29	0.25	0.21	0.15	0.11	0.58***	1.00	
9 MCAR	0.51***	0.11	0.56***	0.03	0.55***	0.36*	0.32*	0.34*	1.00

* $p < .05$ ** $p < .01$ *** $p < .001$

EWRC – English WRC, E. Acc – English Accuracy, A1LV – German reading comprehension, A1HV – German listening comprehension, A1SK – German written communication, A1MK – German oral communication, GWRC – German WRC, G Acc – German Accuracy, MCAR – Minnesota Comprehensive Assessment – Reading.

English ORF is moderately correlated with MCA Reading results ($r = 0.51$, $p < 0.00$), and German ORF and MCA Reading results have an even smaller correlation ($r = 0.32$, $p < 0.05$). German and English ORF are still strongly correlated ($r = 0.74$, $p < 0.00$). German reading comprehension is moderately correlated with MCA Reading results ($r = 0.56$, $p < 0.01$). However, German listening comprehension is not significantly correlated with MCA Reading results.

The smaller sample size and weaker correlations among variables align with the step-wise linear regression results for the fifth grade. The final regression model (Table 14) explains even less variability in MCA Reading scores than the fourth-grade model ($R^2 = 0.53$). German ORF is never a significant predictor in any of the models, even before English ORF is added. The addition of the significant predictor English ORF resulted in an R^2 change of 0.12. These results suggest that although English ORF may be a weak predictor of MCA Reading performance in the fifth-grade immersion setting, a larger sample size or different variables may be needed to explain sufficient MCA Reading performance variability.

Table 7 displays the results of the cross-language analyses predicting to the English outcome measure (MCA Reading). As suggested by the correlation and regression analyses already described, and similar to third and fourth grade results, English ORF in isolation is a significant predictor of MCA Reading scores ($R^2 = 0.27$), although it explains far less variability in the fifth-grade results. German ORF in isolation is also a significant predictor of MCA Reading scores, yet again it describes less overall variability than the English ORF scores and less variability than in the third and fourth grade results ($R^2 = 0.10$). When combined in a single regression model, only the English ORF scores were significant predictors of MCA Reading scores. German ORF does not improve the amount of variability explained.

Table 8 displays the results of the cross-language analyses predicting to the German outcome measure (A1LV). None of the regression analyses yielded significant predictors, nor did any of these models describe a large amount of variability in the German reading comprehension measure. The smaller sample size, along with a skewed outcome variable with limited variability and validity evidence, suggest that these regression results should be interpreted with caution.

Table 14.

Fifth grade step-wise regression model building, predicting end of year English MCA reading assessment.

	Grade 5 Model 1	**Grade 5 Model 2**	**Grade 5 Model 3**	**Grade 5 Model 4**	**Grade 5 Model 5**	**Grade 5 Model 6**
Intercept	565.49 (1.71)	564.04 (61.66)	538.93 (57.77)	495.65 (53.00)	526.43 (59.11)	532.35 (53.37)
Special Education	-5.06 (4.20)	-5.05 (4.28)	-7.14 (4.04)	-5.19 (3.63)	-3.95 (3.76)	-1.18 (3.51)
A1HV	-	0.15 (6.20)	-1.95 (5.79)	2.70 (5.32)	-1.42 (6.33)	-3.28 (5.74)
A1MK	-	-	5.15** (1.87)	2.24 (1.88)	2.35 (1.87)	2.54 (1.69)
A1LV	-	-	-	2.70** (0.81)	2.40** (0.84)	2.07* (0.77)
GWRCW	-	-	-	-	0.11 (0.09)	-0.08 (0.10)
EWRCW	-	-	-	-	-	0.20** (0.07)
R^2	0.04	0.04	0.20	0.38	0.41	0.53

* *p* < .05 ** *p* < .01 *** *p* < .001

EWRC – English WRC, A1LV – German reading comprehension, A1HV – German listening comprehension, A1SK – German written communication, A1MK – German oral communication, GWRC – German WRC.

Diagnostic Accuracy of Reading Fluency in the Immersion Setting

The second purpose of the present study was to evaluate the ability of German and English ORF to accurately identify students who are at risk in reading. The results of the regression analyses previously described suggested that in general, English ORF and German ORF describe some of the same variability in MCA Reading scores. Therefore, English ORF was used as the sole predictor of MCA Reading scores in the following analyses.

Grade 3. Figure 3 displays the diagnostic accuracy results across grades. The results of the third-grade diagnostic accuracy analysis yielded an AUC of 0.91. When the analysis method balanced sensitivity and specificity, sensitivity did not meet criteria (SE = .79), but specificity did (SP = 0.81). In other words, students were not over-identified, but some students at risk of failing the MCA Reading might have been overlooked. The positive predictive power (PPV) of .79 suggests that the "at risk" category on the screener accurately predicts failing. The negative predictive power (NPV) of .81 suggests that the screener category of "not at risk" accurately predicts success on the MCA Reading.

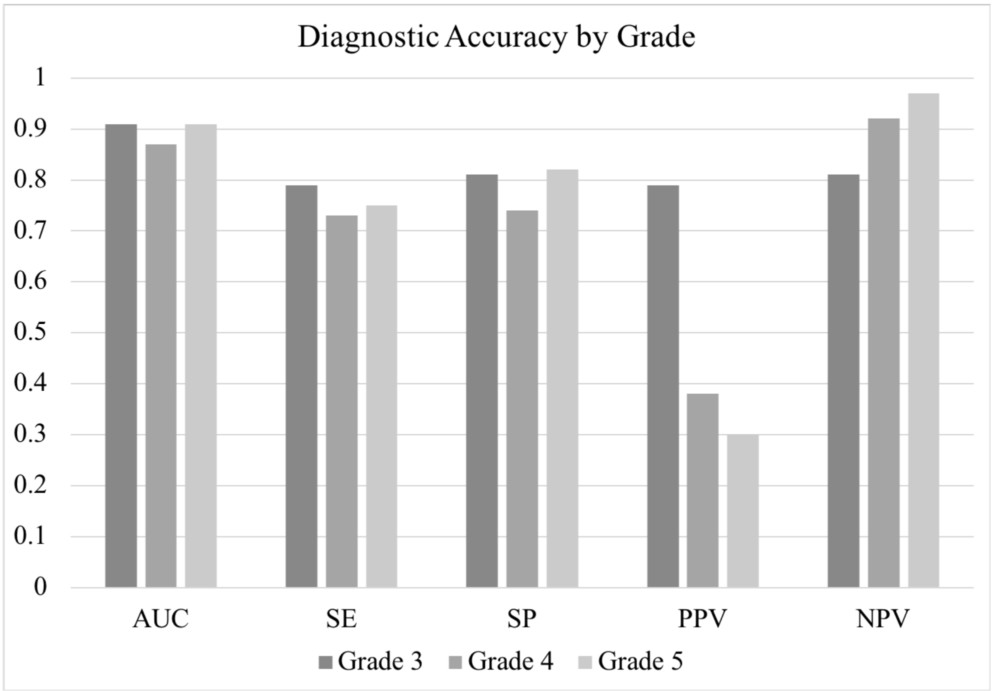

Figure 3. Diagnostic accuracy statistics of English ORF predicting MCA reading scores by grade.

Grade 4. Again, displayed on Figure 3, the diagnostic accuracy results in grade four suggest that English ORF is not as accurate at predicting MCA Reading success in fourth grade. The overall AUC of .87 does not quite meet the .9 standard criterion. Both sensitivity (SE = .73) was below the .8 criterion, and specificity (SP = .74) was slightly above criterion. This suggests that some students were over-identified. The positive predictive power in particular was quite low (PPV = .38), which suggests that the "at risk" status on the screener does not accurately predict failure on the MCA Reading outcome assessment. The "not at risk" status on the screener, however, is much more accurate (NPV = .92). In other words, too many students identified as "at risk" based on the best possible cut-off score on English ORF went on to be successful on the MCA Reading assessment. In resource allocation terms, more students would have been allocated support services than actually needed them.

Grade 5. As in third grade, the overall AUC of .91 meets the criterion. Sensitivity does not quite meet the .8 criterion (SE = .75), which suggests some students who failed the MCA Reading assessment were not identified by the English ORF "at risk" category. Specificity does meet the criterion, however (SP = .82). As in fourth grade, the positive predictive power is quite low (PPV = .30). Again, this suggests that the "at risk" category on the screener was not accurate at predicting failure on the MCA Reading outcome assessment. The "not at risk" status on the screener, similar to fourth grade, is more accurate (NPV = .97). Again, in resource allocation, this suggests that students would have received more support services than actually were needed.

Discussion

The purpose of this study was twofold. First, the study sought to evaluate the extent to which oral reading fluency measures in English and German are predictive of reading outcomes for German immersion emerging biliterates. Second, this study sought to evaluate the ability of German and/or English ORF to accurately identify students who are at risk in reading. In general, English ORF remained the best predictor of both German and English reading outcomes, although decreasing predictive power of English ORF was seen as grades increased. The diagnostic accuracy of English ORF was strongest in third grade and decreased as grades increased.

English and German ORF as Predictors of Reading Outcomes

Theories in second-language literacy acquisition suggest cross-language measurement of reading skills (Bernhard, 2011). Therefore, the use of ORF in English and German as a measure of English and German reading across grades, as well as the relative utility of ORF in German as compared to ORF in English for immersion learners, are reviewed.

German and English ORF as predictors of English reading. In general, English ORF functioned as a predictor of English reading outcomes similarly in the German immersion setting as it has been shown to function in monolingual English settings (Reschly et al., 2009). At grades three, four, and five, English ORF scores demonstrated strong to moderate correlations with MCA Reading scores. German ORF demonstrated weaker correlations with MCA Reading scores. When included in regression analyses predicting MCA Reading performance, German ORF was a significant predictor variable at grades three and four, but only before English ORF was added to the model. English ORF proved the strongest predictor variable across grades.

The present study investigates the use of L1 ORF (English) to predict English outcomes for third, fourth, and fifth graders who began English instruction in third grade.

Outcomes for English speakers in early, total immersion programs demonstrated an initial lag in English literacy. However, the eventual English literacy performance was similar to that of monolingual English peers within one to two years of the start of English instruction (Genesee & Jared, 2008). In the present study, we would expect students reading scores' to be below expectations for monolingual English speakers in third grade and fourth grade and have caught up to monolingual English speaking students by about fifth grade. In other words, it would be expected that English ORF would function similarly to how it functions for monolingual English speakers in fifth grade, but not in third and fourth grade. As noted by previous researchers (Genesee & Jared, 2008), it is possible that the resource-rich home environment provides students included in the present study more English literacy exposure than other immersion students may receive.

Predictive properties of ORF across grades. Overall, English ORF remained a significant predictor across all grades. Across grades three, four, and five there was a trend of decreasing correlations between English ORF and MCA Reading scores, as well as decreasing correlations between German ORF and MCA Reading scores. Similarly, the amount of variability explained by either a full model (with German language proficiency variables) or a simple model (with only German and English ORF) also decreased across grades three, four, and five. This pattern of decreasing explanatory power as grades increase mirrors the pattern seen in research on ORF with monolingual English speakers (Silberglitt et al., 2006), as well as the results of the synthesis in study one of the present project. Results from the present study are somewhat surprising given that L1 reading of immersion students typically reaches parity with the reading of monolingual peers in fifth grade (Lindholm-Leary & Genesee, 2014). It is also important to note that in the present study, the sample of fifth grade students was small, so cross-grade comparisons should be made with caution.

German language proficiency as a mediator between English ORF and English reading. The inclusion of German language proficiency variables in regression analyses at the third grade did not significantly improve the overall variability explained. Furthermore, German oral communication and listening comprehension were not significant predictors of MCA scores in the third-grade regression analyses. Of the German language variables, only German reading comprehension was a significant predictor of MCA Reading scores in regression analyses with third grade students' scores. The results of the regression models conducted with fifth grade students' scores suggested, similar to third grade results, that German reading comprehension and English ORF were significant predictors of MCA reading scores. No other German language proficiency variables were significant predictors of MCA scores. It is possible that the psychometric properties of the German language proficiency assessment were insufficient to be accurate measures of language proficiency for the present study.

As with the results in third grade, fourth grade German reading comprehension was a significant predictor of MCA Reading scores in regression analyses. However, the results of the regression models conducted with fourth grade students' scores suggested that listening comprehension was an important predictor of MCA scores, even after German reading comprehension, German ORF, and English ORF were added to the model. In the German immersion school, students in fourth grade still receive all core content in German, except for one English class. In fifth grade, students begin receiving core instruction like Math in English. In other words, fourth grade demands the greatest amount of time listening in German. This might explain why German listening comprehension is a significant predictor only in the fourth grade regression model results.

Previous research (Bialystok, 2007) suggests that some skills that may be part of reading comprehension, such as metalinguistic awareness, may transfer across languages. This transfer of skills might explain the moderate correlations between MCA Reading scores and the presence of German reading comprehension as a significant predictor in regression models across grades three, four, and five. In other words, some of the skills that students use to be successful on the MCA Reading assessment and some of the skills that students use to be successful on the measure of German reading comprehension may be the same across languages. From the present results, it isn’t clear whether it is metalinguistic awareness, comprehension strategies, or some other underlying skill or ability that explains the correlation between reading measures across languages.

Relative explanatory power of German ORF as compared to English ORF. For analyses across grades three, four, and five, German ORF did not predict variability in MCA Reading scores after English ORF had been added to the model. This suggests that for students in the German immersion environment, German ORF predicts English outcomes only via a possible mediator of English ORF. This is similar to the pattern found in previous research with ELLs in a dual-language environment (Baker, Park, & Baker, 2012): ORF scores function as predictors of reading outcomes within languages best. Previous research suggests that word-level or decoding skills may transfer more readily across languages (Comeau et al., 1999; Lee & Schallert, 1997; Verhoeven, 1994); it is possible that relationships between ORF scores and reading outcome measures across languages are better described by language proficiency or vocabulary in the second language.

Dissimilar to previous research results is English ORF's relationship with the German reading comprehension measure. In isolation, both German and English ORF were significant predictors of German reading comprehension scores at third and fourth grade. However, once German ORF was added to the model in addition to English ORF, both predictors were non-significant. This suggests that rather than a mediation relationship, German ORF and English ORF describe the same variability in the German reading comprehension outcome. These dissimilar results may be a result of a different population of student receiving a different type of instruction. Previous research was conducted with Spanish-speaking ELLs in the US, learning a majority language. The present research was conducted with English-speaking German language learners in the US, learning a foreign language. Differences in language exposure and type of instruction may also be related to differences in vocabulary and language proficiency. Finally, it is important to note that the relatively low variability in the German comprehension outcome variable limit the strength of conclusions that can be made about the relationship between German ORF, English ORF, and German reading comprehension for the immersion learners.

The Accuracy of ORF as a Screener in the German Immersion Setting

This study sought to evaluate the ability of German and/or English ORF to accurately identify students who are at risk in reading. Specifically, how accurately can either German ORF, English ORF, or a combination of the two predict which students will not meet proficiency on the MCA?

The strongest predictor of MCA proficiency was English ORF. Therefore, English ORF was the only predictor used in diagnostic accuracy analyses. The overall AUC in each grade was quite high, although fourth grade AUC did not meet criterion. Third grade results suggested that English ORF is an acceptably accurate screening tool, with most coefficients meeting or close to meeting criteria. Fourth and fifth grade results suggested that a score of "at risk" on the English ORF screener may not mean a student is truly at risk. This could result in allocation of more resources than is necessary to ensure all students pass the MCA Reading assessment.

Previous research would suggest that for emerging biliterates, cut scores may need to be adjusted locally to be more accurate (cf. Hosp, Hosp, & Dole). The research on ORF cut scores would align with research (Lindholm-Leary & Genesee, 2014) that suggests that immersion students' reading skills may have a different trajectory than the reading skills of monolingual English students. In the present study, cut scores derived from diagnostic accuracy analyses were lower than publisher provided cut scores. At third grade, the cut score provided by analysis was 112, whereas the benchmark under which students are considered at "some risk" in the FastBridge system was 116. The cut score at the fourth grade provided by analysis was 128, whereas the publisher-provided cut score was 136. Finally, the fifth grade cut score provided by analysis was 138, and the publisher provided score was 151. These results provide support for the recommendation of local norms and benchmarks in dual-language environments.

Limitations

As with other research on emerging biliterates, the specific dual-language setting is an important variable to consider. The generalizability of the present results is limited by the type of language spoken, the type of language instruction provided, as well as by the largely homogeneous student population at the school.

The outcome measures used in the present study are also a limitation. State accountability assessments like the MCA Reading assessment measure several variables related to reading, including comprehension and vocabulary. Furthermore, the German language proficiency measures in the present study are not as psychometrically sound as other language proficiency measures might be. Also, the lack of variability in the German reading comprehension scores limited the analysis possible with the present dataset. The German reading comprehension variable was not usable as a dependent variable in a regression analysis with multiple predictors.

The German ORF measures were developed for the present German immersion school setting. These measures were developed to mirror the development of the English ORF measures, in that the content was primarily narrative. The narrative stories likely required conversational language proficiency in German, rather than academic language proficiency in German. Because German is the language of instruction, students may be using academic German more than they might use conversational German (Fortune, 2001). Therefore, it would be worthwhile in future studies to validate different types of passages similar to the work that has been done in English ORF validation studies.

Finally, base rates are important to consider as part of diagnostic accuracy analyses. In the present study, 52% of third grade students, 82% of fourth grade students, and 92% of fifth grade students were proficient on the MCA Reading assessment. The decreasing positive predictive power across grades three through five could be a result of differences in base rate, rather than differences in the utility of ORF as a reading screening tool.

Future Directions

To date, research in language immersion students' reading skills has focused largely on describing the variables relevant to reading outcomes in the immersion setting (cf. Comeau et al., 1999). However, there remain relatively few screening tools that are validated for the immersion environment (Fortune, with Menke, 2010; Genesee & Fortune, 2014). Therefore, future work should focus on building validity evidence for reading screening tools.

Results of the present study suggest that local variables, such as the amount of time spent in different languages of instruction, could influence the functioning of ORF as a reading screener. Therefore, future work should include local variables in the analysis. These variables include time in instruction in either language, content of instruction, language background of students in the school, and the relationship between first and second languages spoken.

Furthermore, the results of the present study suggest that local norms may be important to develop when validating reading screening tools in immersion environments. Future research could systematically investigate the functioning of ORF in a variety of immersion settings and build methods for local immersion or dual-language schools to easily create their own norms for screening tools.

Conclusion

To ensure early support and intervention, screening for reading difficulties should occur for every student. However, little is known about the functioning of commonly used reading screening tools in immersion environments. The present study suggests that for students in an early total foreign language immersion environment, ORF may be a reasonable tool to identify students at risk for not passing the state accountability assessment.

CHAPTER 4: General Discussion

There are increasingly diverse populations of emerging biliterates in our schools. Despite the politicization of English Language Learners and dual-language programs (Christian, 2011), we must continue to build empirical evidence about the education of emerging biliterates in order to ensure equitable educational opportunities for all students. One type of needed empirical evidence is related to the use of screening tools to identify students at risk of poor learning outcomes.

The present project sought to build evidence about the use of ORF as a reading screening tool with emerging biliterates. First, through study one, the project synthesized available empirical evidence about the functioning of ORF as a measure of reading, and the accuracy of ORF in the identification of emerging biliterates at risk for poor reading outcomes. Results suggested that whereas correlations between ORF and reading outcomes were moderate to strong, diagnostic accuracy generally did not meet criteria. Next, through study two, the project conducted a small study in an immersion environment to build evidence related to the use of ORF within the increasing number of immersion schools in the US. Results to study two suggested that for third, fourth, and fifth graders in an early total foreign immersion program, ORF in students' first language may be a good and reasonably accurate predictor of L1 reading outcomes.

Emerging Biliterates as a Heterogeneous Population

To continue to build empirical evidence about the use of ORF as a screening tool with emerging biliterates, researchers must recognize that emerging biliterates are not a homogeneous population. Type of instruction (dual language, English-only support, etc.), language of instruction, first and second (and other) language, and relative language proficiency are facets that describe different groups within the category of emerging biliterates.

In study one, the synthesized studies included were primarily conducted with ELLs in the United States. However, many studies failed to describe the type of instruction received by participants. Results from multiple longitudinal studies suggest that the language and type of instruction received can impact language proficiency and reading proficiency outcomes for students learning to read in more than one language (Lindholm-Leary & Genesee, 2014; Umansky, Valentino, & Reardon, 2016). Indeed, results of study two of the present project, conducted in an immersion setting, found results that differed from similar studies with ELLs in dual-language programming (Baker, Park, Baker, 2012). This suggests that future empirical studies must thoroughly describe the type of instructional programming provided to emerging biliterates.

Similarly, many studies synthesized in study one failed to sufficiently describe the language background of participants. Several of these studies collapsed students with a variety of first languages into one "ELL" group. Biliteracy evidence and theory both suggest that the similarity between languages may change the functioning of cross-language compensatory processes that occur when reading in a second language (Koda, 2008). In fact, many of the studies that were excluded from the synthesis in study one used the demographic flag of "ELL" as a control variable in larger analyses about the functioning of ORF as a screening tool. Results of study two, conducted with English-language speakers in the US learning German, are necessarily limited in generalizability to majority language speakers in early total immersion programs. The generalizability may even be limited to majority language speakers in early total immersion programs, wherein the target language is orthographically shallow and similar to the first language.

Emerging Biliterates and Reading Theory

Results of the synthesis in study one suggest that the relationship between ORF and reading outcomes may be influenced by other variables, such as language proficiency, vocabulary, and oral language proficiency in the second language. Study two empirical results found that reading comprehension across languages, as well as listening comprehension in L2, may be important variables to consider when investigating the relationship between ORF and reading outcomes. Biliteracy theories would support these conclusions; variability in second-language reading comprehension is likely related to literacy in the first language as well as language knowledge (including vocabulary and oral language). Taken together, this suggests that the conclusion that ORF functions similarly for ELLs as it does for monolingual English-speaking students is premature.

It is also important to note that much of the research conducted about the reading comprehension of biliterates has been conducted with adults (Bernhardt, 2011). Although evidence has been gathered across different levels of second-language proficiency, this work may not generalize to young emerging biliterates. Just as reading theory developed based on monolingual English readers may not completely apply to biliteracy theory (Bernhardt, 2005), application of adult biliteracy models cannot be applied to young emerging biliterates (Bilaystok, 2007). Young students who are simultaneously learning a language and learning to read represent a different population than adults who have mastered literacy skills in at least one language before attempting them in another (Paradis, Genesee, & Crago, 2011). Therefore, continued research about young emerging biliterates' reading comprehension would address a gap in the present literature.

Emerging Biliterates in Schools Now

Despite the relatively small amount of empirical evidence to support the use of ORF as a screening tool with emerging biliterates, millions of emerging biliterates are in schools in the United States today and should be screened. Therefore, it is necessary to provide the best possible practical solutions, while considering the limitations of the available empirical evidence.

First, reading screening tools should be validated locally whenever possible. As with other educational innovations or instruction, new assessment methods should be treated as hypotheses in an experiment. The use of ORF as a reading screening tool should be considered a new, un-validated method. Therefore, schools and districts with specific populations of students should investigate the functioning of ORF for their students. For example, a school with a large number of ELLs could investigate whether the "at-risk" category on a published ORF measure accurately predicts not passing a state accountability measure. This is a relatively straightforward table of correct vs. incorrect classification, and it does not require sophisticated statistical methods or knowledge. This table could be created separately for ELLs with different language backgrounds, levels of language proficiency, and receiving different types of language instruction. Over time, it may become clear whether ORF is functioning as an accurate reading screening tool or not.

Next, should accurate screening not be achieved with ORF alone, multiple-gate procedures might be applied. In this procedure, ORF could be administered to the population of emerging biliterates first. All students who may possibly be at risk could then be given a second measure, such as vocabulary or oral language. As with ORF alone, this could be tested with simple tables of accurate identification of students.

Finally, it is important that schools continue to evaluate whether instruction and educational programs are effective for their emerging biliterates. For example, longitudinal analysis of outcomes in language proficiency and reading would be helpful to understand the benefits of certain types of dual-language programs over others (cf. Umansky, Vellutino, & Reardon, 2016). Again, educational programming for emerging biliterates, even mandated English-only programming, should be regularly and thoughtfully evaluated.

Conclusion

The present study sought to provide necessary evidence about the use of ORF as a screening tool with emerging biliterates. ORF in L1 appears to function well for majority language speakers learning a foreign language in an early total immersion environment. However, despite claims about the utility of ORF for ELLs (Sandberg & Reschly, 2011), ORF in English for ELLs does not appear to accurately identify students at risk for poor reading outcomes. Further research is needed to learn more about the variables related to ORF functioning for ELLs, as well as develop methods to improve the accuracy of ORF as a screening tool for ELLs. Based on the results of the present project, educators should treat ORF as a screening tool with emerging biliterates as a novel educational practice that still requires validation.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Ardoin, S. P., Suldo, S. M., Witt, J., Aldrich, S., & McDonald, E. (2005). Accuracy of Readability Estimates' Predictions of CBM Performance. *School Psychology Quarterly, 20*(1), 1.

Arizona Department of Education. (2007). *Arizona English Language Learner Assessment (AELLA): Technical Manual*.

Arizona Department of Education. (2009). *Arizona's Instrument to Measure Standards: 2007 Technical Report*.

August, D., McCardle, P., & Shanahan, T. (2014). Developing literacy in English language learners: Findings from a review of the experimental research. *School Psychology Review*, *43*(4), 490-498.

Baker, D. M. L. (2007). *Relation between oral reading fluency and reading comprehension for Spanish-speaking students learning to read in English and Spanish*. (Unpublished doctoral dissertation). University of Oregon, Eugene, Oregon.

Baker, S. K., & Good, R. H. (1995). Curriculum-based measurement of English reading with bilingual Hispanic students: A validation study with second-grade students. *School Psychology Review, 24* (4), 561-578.

Baker, D. L., Park, Y., & Baker, S. K. (2012). The reading performance of English learners in grades 1–3: the role of initial status and growth on reading fluency in Spanish and English. *Reading and Writing, 25*(1), 251-281.

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics, 25*, 133-150.

Bernhardt, E. (2011). *Understanding advanced second-language reading*. New York, NY: Routledge.

Betts, J., Muyskens, P., & Marston, D. (2006). Tracking the progress of students whose first language is not English towards English proficiency: Using CBM with English language learners. *MinneTESOL/WITESOL Journal, 23*, 15–37.

Bialystok, E. (1988). Levels of bilingualism and levels of linguistic awareness. *Developmental Psychology, 24*(4), 560.

Bialystok, E. (2001). *Bilingualism in development: Language, literacy, and cognition*. New York, NY: Cambridge University Press.

Bialystok, E. (2007). Acquisition of literacy in bilingual children: A framework for research. *Language Learning, 57*, 45-77.

Bialystok, E. (2011). Reshaping the mind: The benefits of bilingualism. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 65*(4), 229.

Bialystok, E., Luk, G., & Kwan, E. (2005). Bilingualism, biliteracy, and learning to read: Interactions among languages and writing systems. *Scientific Studies of Reading, 9*(1), 43-61.

Björnsson, C. H. (1968). *Lesbarkeit durch Lix*. Pedagogiskt centrum, Stockholms skolförvaltn.

Bournot-Trites, M. (2008). Fostering reading acquisition in French immersion. In *Encyclopedia of Language and Literacy Development* (pp. 1-8). London, ON: Canadian Language and Literacy Research Network.

Bruck, M., Genesee, F., & Caravolas, M. (1997). A cross-linguistic study of early literacy acquisition. In B.A. Blachman (Ed.), *Foundations of reading acquisition and dyslexia: Implications for early intervention* (pp. 145-162). Mahwah, N.J.: Lawrence Erlbaum Associates.

California Department of Education. (2007). *Technical report for the California English Language Development Test (CELDT)*. Monterey: Author.

Champely, S. (2016). pwr: Basic Functions for Power Analysis. R package version 1.2-0. https://CRAN.R-project.org/package=pwr

Center for Applied Linguistics (2011). Directory of Foreign Language Immersion Programs in U.S. Schools. Retrieved February 2, 2016 from http://webapp.cal.org/Immersion/Doc/Growth%20of%20Language%20Immersion%20Programs%20in%20the%20US%201971-2011.pdf

Christ, T. J., Van Norman, E. R., & Nelson, P. M. (2016). Foundations of Fluency-Based Assessments in Behavioral and Psychometric Paradigms. In *The Fluency Construct* (pp. 143-163). Springer New York.

Christian, D. (2011). Dual language education. In E. Hinkel (Ed.) *Handbook of research in second language teaching and learning Volume II* (pp. 3-20). New York: Routledge.

Chu, S. Y., & Flores, S. (2011). Assessment of English language learners with learning disabilities. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 84*(6), 244-248.

CTB McGraw-Hill (2001). TerraNova. Second Edition. Monterey, CA: Author.

CTB McGraw-Hill (2014). *Wisconsin Knowledge and Concepts Examinations Fall 2013 WKCE Technical Report.* Monterey, CA: Author.

Comeau, L., Cormier, P., Grandmaison, É., & Lacroix, D. (1999). A longitudinal study of phonological processing skills in children learning to read in a second language. *Journal of Educational Psychology, 91*(1), 29-43.

Cummins, J. (1980). The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue. *Tesol Quarterly*, 175-187.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In *Schooling and language minority students: A theoretical framework*, 3-50. Los Angeles, CA: Evaluation, Dissemination, and Assessment Center, California State University.

Cummings, K. D., Baker, D. L., & Good, R. H. (2006). Guía en inglés para la administración y calificación de IDEL. In D. L. Baker, R. H. Good, N. Knutson, & J. M. Watson (Eds.), Indicadores Dinámicos del Éxito en la Lectura (7a ed.). Eugene, OR: Dynamic Measurement Group.

Crosson, A. C., & Lesaux, N. K. (2010). Revisiting assumptions about the relationship of fluent reading to comprehension: Spanish speakers' text-reading fluency in English. *Reading and Writing,* 23, 475 – 494. doi:10.1007/s11145-009-9168-8

Dalton, S. (1979). Validation of the Language Assessment Scales. *Educational and Psychological Measurement, 39*(4), 1001-1003.

DeAvila, E., & Duncan, S. (1977). Language Assessment Scales (2nd ed.). Corte Madera, CA: Linguametrics Group.

Da Fontoura, H. A., & Siegel, L. S. (1995). Reading, syntactic, and working memory skills of bilingual Portuguese-English Canadian children. *Reading and Writing, 7(*1), 139-153.

de Valenzuela, J. S., & Niccolai, S. L. (2004). Language development in culturally and linguistically diverse students with special education needs. In L.M Baca & H.T. Cervantes (Eds.) *The bilingual special education interface*, 125-161.Upper Saddle River, NJ: Pearson.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional children, 52*(3), 219-232.

Deno, S. L. (1990). Individual Differences and Individual Difference: The Essential Difference of Special Education. *The Journal of Special Education, 24*(2), 160–173. http://doi.org/10.1177/002246699002400205

Deno, S. L. (2003). Developments in Curriculum-Based Measurement. *The Journal of Special Education, 37*(3), 184–192. http://doi.org/10.1177/00224669030370030801

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

Denton, C. A., Barth, A. E., Fletcher, J. M., Wexler, J., Vaughn, S., Cirino, P. T., ... & Francis, D. J. (2011). The relations among oral and silent reading fluency and comprehension in middle school: Implications for identification and instruction of students with reading difficulties. *Scientific Studies of Reading, 15*(2), 109-135.

Dixon, L. Q., Zhao, J., Shin, J. Y., Wu, S., Su, J. H., Burgess-Brigham, R., ... & Snow, C. (2012). What We Know About Second Language Acquisition A Synthesis From Four Perspectives. *Review of Educational Research, 82*(1), 5-60.

Dunn, L. M., & Dunn, D. M. (2007). Peabody picture vocabulary test (4th edition). Minneapolis, MN: Pearson Assessments.

Dunn, L. M., Padilla, E. R., Lugo, D. E., & Dunn, L. M. (1986). Test de Vocabulario en Imágenes Peabody: Adaptación Hispanoamericana. Circle Pines, MN: American Guidance Service.

Durgunoglu, A. Y. (2002). Cross-linguistic transfer in literacy development and implications for language learners. *Annals of Dyslexia*, 52,189–204. doi:10.1007/s11881-002-0012-y

Francis, Lesaux, & August, (2006). Language of instruction. In D. August & T. Shanahan (Eds.), *Developing literacy in second language learners. Report of the National Reading Panel on language minority and youth*, (365-414). Mahwah, NJ: Lawrence Erlbaum Associates.

Echols, J. M. Y. (2010). *The Utility of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in Predicting Reading Achievement.* (Unpublished doctoral dissertation). Seattle Pacific University, Seattle, Washington.

Educational Testing Service. (2007). California Standards Tests (CST) technical report: Spring 2006 administration. Princeton, NJ: Author.

Erdos, C., Genesee, F., Savage, R., & Haigh, C.A. (2010). Individual differences in second language reading outcomes. *International Journal of Bilingualism, 1*5(1), 3-25.

Farmer, E. (2013). *Examining predictive validity and rates of growth in curriculum-based measurement with English language learners in the intermediate grades.* (Unpublished doctoral dissertation). Loyola University, Chicago, Illinois.

Felt, J.N. (2015). *Word calling on curriculum based measures in English language learners.* (Unpublished Doctoral dissertation). University of Wisconsin, Madison, Wisconsin.

Fortune, T. W. with Menke, M. R. (2010). *Struggling learners & language immersion education: Research-based, practitioner-informed responses to educators' top questions* (CARLA Publication Series). Minneapolis, MN: University of Minnesota, the Center for Advanced Research on Language Acquisition.

Fortune, T. W. (2001). *Understanding immersion students: Oral language use a mediator of social interaction in the classroom*. (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.

Fortune, T. W. (2011). Struggling learners and the language immersion classroom. In D. J. Tedick, D. Christian, & T.W. Fortune (Eds.), *Immersion education: Practices, policies, possibilities*, (251-270). Buffalo, NY: Multilingual Matters.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*(2), 188-193.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to Response to Intervention: What, Why, and How Valid Is It? *Reading Research Quarterly*, 41, 93–99. doi:10.1598/RRQ.41.1.4

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*, 7–21.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific studies of reading, 5*(3), 239-256.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20-28.

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

Fuchs, L. S., & Vaughn, S. (2012). Responsiveness-to-intervention: A decade later. *Journal of learning disabilities, 45*(3), 195-203.

Ganan, B. J. (2012). *The Fluidez en La Lectura Oral (FLO) Portion of the Indicadores Dinamicos De Exito en La Lectura (IDEL) and the English language portion of the Illinois Standard Achievement Test (ISAT): A correlational study of second and third grade English language learners*. (Unpublished doctoral dissertation). Roosevelt University, Schaumburg, Illinois.

García, G. E., McKoon, G., & August, D. (2006). Synthesis: Language and literacy assessment. In D. August & T. Shanahan (Eds.), *Developing literacy in second language learners. Report of the National Reading Panel on language minority and youth,* 583-696. Mahwah, NJ: Lawrence Erlbaum Associates.

Genesee, F. (1985). The Linguistic and Academic Development of English-Speaking Children in French Schools: Grade 4 Outcomes. *Canadian Modern Language Review, 41(*4), 669-85.

Genesee, F. (1987). *Learning through two languages: Studies of immersion and bilingual education.* Newbury house publishers.

Genesee, F., & Fortune, T. W. (2014). Bilingual education and at-risk students. *Journal of Immersion and Content-Based Language Education, 2*(2), 196-209.

Genesee, F., & Jared, D. (2008). Literacy development in early French immersion programs. *Canadian Psychology, 49*, 140–147. doi:10.1037/0708-5591.49.2.140

Genesee, F., Savage, R., Erdos, C., & Haigh. C. (2013). Identification of reading difficulties in students schooled in a second language. In V. Gathercole (Ed.), *Solutions for the Assessment of Bilinguals* (pp 10-35). Buffalo, NY: Multilingual Matters.

Genesee, F., Tucker, G. R., & Lambert, W. E. (1975). Communication skills of bilingual children. *Child development*, 1010-1014.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional children, 71*(2), 149-164.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology*, *45*(2), 117-135.

Good, R. H., & Kaminski, R. A. (2011). Dynamic Indicators of Basic Early Literacy Skills Next. Eugene, OR: Dynamic Measurement Group. Retrieved from http://www.dibels.org/

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and special education, 7*(1), 6-10.

Grasparil, T. A., & Hernandez, D. A. (2015). Predictors of Latino English Learners' Reading Comprehension Proficiency. *Journal of Educational Research and Practice, 5*(1).

Grosjean, F. (2010). *Bilingual*. Harvard University Press.

Hakuta, K., Butler, Y.G., & Witt, D. (2000). How long does it take English learners to attain proficiency? The University of California Linguistic Minority Research Institute Policy Report 2000-1.

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

Harcourt Brace & Company (1997b). Stanford Achievement Test Series – Ninth Edition: Technical data report. San Antonio, TX: Harcourt Brace & Company.

Harcourt Brace (2003). Stanford achievement test—tenth edition: Technical data report. San Antonio, TX: Author.

Harcourt Educational Measurement (2005). Aprenda: La prueba de logros en español, tercera edición. Technical Manual. San Antonio, TX.

Hosp, M. K., & Fuchs, L. S. (2005). Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Review, 34*, 9–26.

Hosp, J. L., Hosp, M. A., & Dole, J. K. (2011). Potential bias in predictive validity of universal screening measures across disaggregation subgroups. *School Psychology Review, 40*(1), 108.

Howard, E. R., Sugarman, J., Christian, D., Lindholm-Leary, K. J., & Rogers, D. (2007). Guiding principles for dual language education (2nd ed.). Washington, DC: Center for Applied Linguistics. Retrieved from http://www.cal.org/twi/pdfs/guiding-principles.pdf

Ilinois State Board of Education (2010a). Interpretive Guide. Illinois Standards Achievement Test 2010. Retrieved https://www.isbe.net/Documents/ISAT_Interpr_Guide_2010.pdf#search=ISAT%20interpretive%20guide

Jared, D., Cormier, P., Levy, B. A., & Wade-Woolley, L. (2011). Early predictors of biliteracy development in children in French immersion: A 4-year longitudinal study. *Journal of educational psychology, 103*(1), 119.

Jimerson, S. R. (2000). ORAL-J: The administration and technical manual. Available from S. R. Jimerson, University of California, Santa Barbara, California, 93106-9490.

Jimerson, S. R., Hong, S., Stage, S., & Gerber, M. (2013). Examining oral reading fluency trajectories among English language learners and English speaking students. *Journal of New Approaches in Educational Research, 2*(1), 3-11.

Jitendra, A. K., Burgess, C., & Gajria, M. (2011). Cognitive strategy instruction for improving expository text comprehension of students with learning disabilities: The quality of evidence. *Exceptional Children, 77*(2), 135-159.

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.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review, 42*(4), 448.

Karlsen, B., & Gardner, E. (1985). Stanford Diagnostic Reading Test (3rd ed.). San Antonio, TX: The Psychological Corporation.

Kim, Y. S. (2012a). The relations among L1 (Spanish) literacy skills, L2 (English) language, L2 text reading fluency, and L2 reading comprehension for Spanish-speaking ELL first grade students. *Learning and Individual Differences, 22*(6), 690-700.

Kim, J. S. (2012b). Evaluating the Predictive Validity of DIBELS Literacy Measures with Third Grade Spanish-Speaking English Language Learners. (Unpublished doctoral dissertation). University of California, Riverside, California.

Klare, George R. Readability. In Handbook of Reading Research, edited by P. David Pearson. New York, NY: Longmans, 1984.

Knight-Teague, K., Vanderwood, M. L., & Knight, E. (2014). Empirical investigation of word callers who are English learners. *School Psychology Review, 43*(1), 3.

Koda, K. (2002). Writing systems and learning to read in a second language. In L. Wenling, J.S. Gaffney, & J. L Packard (Eds.), *Chinese children's reading acquisition* (pp. 225-248). New York, NY: Springer Science.

Koda, K. (2008). Impacts of prior literacy experience on second language learning to read. In K. Koda & A. Zehler (Eds.) *Learning to read across languages: Cross-linguistic relationships in first-and second-language literacy development* (pp.68-96). New York, NY: Routledge.

Kranzler, J. H., Brownell, M. T., & Miller, M. D. (1998). The construct validity of curriculum-based measurement of reading: An empirical test of a plausible rival hypothesis. *Journal of School Psychology, 36*, 399– 415.

Krashen, S. D. (1981). Bilingual education and second language acquisition theory. In Schooling and language minority students: A theoretical framework, 51-79. Los Angeles, CA: Evaluation, Dissemination, and Assessment Center, California State University.

Lee, J. W., & Schallert, D. L. (1997). The relative contribution of L2 language proficiency and L1 reading ability to L2 reading performance: A test of the threshold hypothesis in an EFL context. *Tesol Quarterly*, *31*(4), 713-739.

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 Reading Panel on language minority and youth,* 583-696. Mahwah, NJ: Lawrence Erlbaum Associates.

Lindholm-Leary, K., & Genesee, F. (2014). Student outcomes in one-way, two-way, and indigenous language immersion education. *Journal of Immersion and Content-Based Language Education, 2*(2), 165-180. doi 10.1075/jicb.2.2.01linissn 2212–8433

Lyster, R. & Tedick, D. (2014). Research perspectives on immersion pedagogy: Looking back and looking forward. *Journal of Immersion and Content-Based Language Education, 2*:2, 210-224. doi10.1075/jicb.2.2.04lys

MacCoubrey, S. J., Wade-Woolley, L., Klinger, D., & Kirby, J. R. (2004). Early identification of at-risk L2 readers. *Canadian Modern Language Review, 61*, 11–28. doi:10.3138/cmlr.61.1.11

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). Gates MacGinitie Reading Tests. Itasca, IL: Riverside Publishing.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dreyer, L. G., & Hughes, K. E. (2009). Gates MacGinitie Reading Tests, Fourth Edition. Rolling Meadows, IL: Riverside.

Mancilla-Martinez, J., & Lesaux, N. K. (2011). The gap between Spanish speakers' word reading and word knowledge: A longitudinal study. *Child development, 82*(5), 1544-1560.

McGraw-Hill. (2003). TerraNova second edition: California Achievement Tests technical report. Monterey, CA: Author.

Meisinger, E. B., Bradley, B. A., Schwanenflugel, P. J., & Kuhn, M. R. (2010). Teachers' perceptions of word callers and related literacy concepts. *School Psychology Review*, *39*(1), 54.

Meisinger, E. B., Bradley, B. A., Schwanenflugel, P. J., Kuhn, M. R., & Morris, R. D. (2009). Myth and reality of the word caller: The relation between teacher nominations and prevalence among elementary school children. *School Psychology Quarterly*, *24*(3), 147.

Mellard, D., McKnight, M., & Jordan, J. (2010). RTI tier structures and instructional intensity. *Learning Disabilities Research & Practice, 25*(4), 217-225.

Millett, J. R. (2011). *The Ability of Oral Fluency to Predict Reading Comprehension Among ELL Children Learning to Read* (Unpublished doctoral dissertation). Arizona State University.

Minnesota Department of Education (2015). Technical manual for Minnesota Title I and Title III Assessments. Retrieved from http://education.state.mn.us/MDE/dse/test/mn/Tech/

Minnesota Department of Education (2017). Yearbook tables for Minnesota Title I and Title III Assessments for the academic year 2015-2016. Retrieved from http://education.state.mn.us/MDE/dse/test/mn/Tech/

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

Myers, M. L. (2009). *Achievement of children identified with special needs in two-way Spanish immersion programs* (Unpublished doctoral dissertation). The George Washington University.

Nam, J. E. (2012). *An examination of the predictive validity of early literacy measures for Korean English language learners.* (Unpublished doctoral dissertation). University of California, Riverside, California.

National Center for Education Statistics. (2016). The condition of education. Retrieved February 2, 2016, from https://nces.ed.gov/programs/coe/indicator_cgf.asp

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

Northwest Evaluation Association. (2002). Northwest Achievement Levels Test. Lake Oswego, OR: Northwest Evaluation Association.

Northwest Evaluation Association (2015). Measures of academic progress. Retrieved from www.nwea.org

Oregon Department of Education. (2008). Oregon's statewide assessment system technical report (Vol. 7): Alternate assessment. Salem, OR: Author.

Parker, D. C., Zaslofsky, A. F., Burns, M. K., Kanive, R., Hodgson, J., Scholin, S. E., & Klingbeil, D. A. (2015). A brief report of the diagnostic accuracy of oral reading fluency and reading inventory levels for reading failure risk among second-and third-grade students. *Reading & Writing Quarterly, 31*(1), 56-67.

Paradis, J., Genesee, F., & Crago, M. B. (2011). Dual language development and disorders. *A handbook on bilingualism and second language learning*, *2nd Edition*.

Pasquarella, A., Chen, X., Gottardo, A., & Geva, E. (2015). Cross-language transfer of word reading accuracy and word reading fluency in Spanish-English and Chinese-English bilinguals: Script-universal and script-specific processes. *Journal of Educational Psychology, 107*(1), 96.

Pearson (2007). Washington Assessment of Student Learning Grade 3 (Tech. Rep.). Retrieved March 10th, 2018 http://www.k12.wa.us/assessment/pubdocs/2006Gr3WASLTechReport.pdf

Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-language perspective on speech information rate. *Language, 87*(3), 539-558.

Petzold, A. (2006). Assessment of struggling elementary immersion learners: The St Paul Public Schools Model. The ACIE Newsletter 9, 1-2, 10-11, 13. Retrieved February 2, 2016, from http://carla.umn.edu/immersion/acie/vol9/feb2006_bpractices_strugglinglearners.html

Pearson, Inc. (2012). AIMSweb Technical Manual. Retrieved from http://www.aimsweb.com/

Pretorius, E. J., & Spaull, N. (2016). Exploring relationships between oral reading fluency and reading comprehension amongst English second language readers in South Africa. *Reading and Writing, 29*(7), 1449-1471.

Pro-Ed Inc. (2008) Test of Word Reading Efficiency. Austin, TX: Author.

Quirk, M., & Beem, S. (2012). Examining the relations between reading fluency and reading comprehension for English language learners. *Psychology in the Schools, 49*(6), 539-553.

Ramirez, R. D., & Shapiro, E. S. (2006). Curriculum-based measurement and the evaluation of reading skills of Spanish-speaking English language learners in bilingual education classrooms. *School Psychology Review, 35*(3), 356.

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

Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban, first grade students. *Reading Research Quarterly, 42*, 460 – 466. doi:10.1598/RRQ.42.4.5

Sáez, L., Park, B., Nese, J. F., Jamgochian, E., Lai, C. F., Anderson, D., ... & Tindal, G. (2010). Technical Adequacy of the easyCBM Reading Measures (Grades 3-7), 2009-2010 Version. Technical Report# 1005. Behavioral Research and Teaching.

Sandberg, K. L., & Reschly, A. L. (2011). English learners: Challenges in assessment and the promise of curriculum-based measurement. *Remedial and Special Education, 32(*2), 144-154.

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

Silberglitt, B., Burns, M. K., Madyun, N. I. H., & Lail, K. E. (2006). Relationship of reading fluency assessment data with state accountability test scores: A longitudinal comparison of grade levels. *Psychology in the Schools, 43*(5), 527-535.

Spectrum K12, American Association of School Administrators, Council of Administrators of Special Education, National Association of State Directors of Special Education, & State Title 1 Directors. (2010). Response to intervention (RTI) adoption survey. Retrieved from http://sss.usf.edu/resources/presentations/2010/fasp_summer_inst2010/Resource_SLD/RTI/2010RTIAdoptionSurveyReport.pdf

Stanovich, K. E., & Stanovich, P. J. (1995). How research might inform the debate about early reading acquisition. Journal of Research in Reading, 18, 87–105.

Stokes, N. O. (2010). *Examining the relationship among reading curriculum-based measures, level of language proficiency, and state accountability test scores with middle school Spanish-speaking English language learners*. (Unpublished doctoral dissertation). Loyola University, Chicago, Illinois.

Swain, M. & Lapkin, S. (1982). *Evaluating bilingual education: A Canadian case study.* Clevedon, Eng: Multilingual Matters.

Swets, J. A., R. M. Dawes, & J. Monahan (2000), "Psychological Science Can Improve Diagnostic Decisions", *Psychological Science in the Public Interest 1*: 1–26

Talbott, E., Maggin, D. M., Van Acker, E. Y., & Kumm, S. (2017). Quality indicators for reviews of research in special education. *Exceptionality*, 1-21.

Tedick, D. J., Christian, D., & Fortune, T. W. (2011). The future of immersion education: An invitation to 'dwell in possibility'. In D. J. Tedick, D. Christian, & T.W.

Fortune (Eds.), *Immersion education: Practices, policies, possibilities,* (1-10). Buffalo, NY: Multilingual Matters.

Thompson, L. E., Boyson, B. A., and Rhodes, N.C. (2006). Administrator's Manual for CAL Foreign Language Assessments, Grades K-8. Washington, D.C: Center for Applied Linguistics.

Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of school psychology, 40*(1), 7-26.

TJCC (2015). Formative Assessment System for Teachers: Technical Manual Version 2.0., Minneapolis, MN: Author and FastBridge Learning (www.fastbridge.org)

Umansky, I. M., Valentino, R. A. & Reardon, S. F. (2016). The promise of two-language education. *Helping ELLs Excel 73*(5), 10-17.

US Census Bureau https://www.census.gov/prod/2013pubs/acs-22.pdf

Utah State Office of Education. (2007). Utah ELA CRT technical manual. Salt Lake City, UT. Available at www. Usoe.k12.ut.us.

VanDerHeyden, A. M., Witt, J. C., & Gilbertson, D. (2007). A multi-year evaluation of the effects of a response to intervention (RTI) model on identification of children for special education. *Journal of School Psychology, 45*(2), 225-256.

Vanderwood, M. L., Tung, C. Y., & Checca, C. J. (2014). Predictive validity and accuracy of oral reading fluency for English learners. *Journal of Psychoeducational Assessment, 32*(3), 249-258.

Verhoeven, L. T. (1994). Transfer in bilingual development: The linguistic interdependence hypothesis revisited. *Language learning, 44*(3), 381-415.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes Cambridge, Mass.: Harvard University Press.

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

Wechsler, D. (1999). Wechsler Abbreviated Scale of Intelligence (WASI). San Antonio, TX: Harcourt Assessment.

Weiderholt, J. L., & Bryant, B. R. (2001). Gray Oral Reading Test, Fourth Edition (GORT–4). Austin, TX: PRO-Ed.

WIDA Consortium (2007). Annual technical report no. 2, volume 1 of 3: Description, validity, and student results. Annual technical report for ACCESS for ELLs ® English Language Proficiency Test, Series 101, 2005-2006 administration. Madison, WI: Author.

Wiley, H. I., & Deno, S. L. (2005). Oral reading and maze measures as predictors of success for English learners on a state standards assessment. *Remedial and Special Education, 26*(4), 207-214.

Williams, K.T. (2001). Technical manual: Group Reading Assessment and Diagnostic Evaluation. Circle Pines, MN: American Guidance Service.

Woodcock, R. W. (1987). Woodcock reading mastery test–revised. Circle Pines, MN: American Guidance Service.

Woodcock, R. W. (1991). Woodcock Language Proficiency Battery—Revised. Itasca, IL: Riverside Publishing.

Woodcock, R. W., McGrew, K., & Mather, N. (2001). Woodcock Johnson tests of achievement (3rd edition). Itasca, IL: Riverside Publishing.

Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education, 31(*6), 412-422.

Yovanoff, P., Duesbery, L., Alonzo, J., & Tindal, G. (2005). Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency. *Educational Measurement: Issues and Practice, 24*(3), 4-12.

Zentralstelle für das Auslandsschulwesen (ZFA; 2014). Ausführungsbestimmungen für die Internationalen schulischen Vergleichsarbeiten der Zentralstelle für das Auslandsschulwesen: Auf dem Weg zum DSD – Niveaustufe A1 und Auf dem Weg zum DSD – Niveaustufe A2. Retrieved from http://www.bva.bund.de/DE/Organisation/Abteilungen/Abteilung_ZfA/Auslandsschularbeit/DSD/Vergleichsarbeiten/node.html

Appendix A

Comprehensive description of studies included in systematic synthesis, organized alphabetically within type of validity evidence.

Study	Grade	Sample size	Participants	Type of Instruction	Language Proficiency	ORF measure	Criterion Measure	Validity Evidence
Correlational evidence								
Baker, D. M. L. (2007). *Relation between oral reading fluency and reading comprehension for Spanish-speaking students learning to read in English and Spanish.* (Unpublished doctoral dissertation). University of Oregon.	2	88	L1: Spanish L2: English	Develop. Bilingual: 90 min Spanish 45 min English	unknown	English: DIBELS Spanish: IDEL	English: SAT-10 comprehension Spanish: Aprenda comprehension	DIBELS/SAT10: .75 DIBELS/Aprenda: .54 IDEL/Aprenda: .62 IDEL/SAT10: .55
Baker, S. K., & Good, R. H. (1995). Curriculum-based measurement of English reading with bilingual Hispanic students: A validation study with second-grade students. *School Psychology Review, 24* (4), 561-578.	2	26	L1: Spanish L2: English	ELL 30 min Spanish per day	LAS 14 fluent, 5 limited, 12 minimal	curriculum	SDRT broad reading SDRT reading comprehension	ORF/broad: .53 ORF/comp:.73

Study	Grade	Sample size	Participants	Type of Instruction	Language Proficiency	ORF measure	Criterion Measure	Validity Evidence
Baker, D. L., Park, Y., & Baker, S. K. (2012). The reading performance of English learners in grades 1–3: the role of initial status and growth on reading fluency in Spanish and English. *Reading and Writing, 25*(1), 251-281.	2	156	L1: Spanish L2: English	Develop. bilingual	unknown	English: DIBELS Spanish: IDEL	English: SAT-10 comprehension OAKS broad reading Spanish: Aprenda comprehension	English: Fall to spring 2nd gr.: ORF/comp: .55 3rd gr.: ORF/comp: .56 Spanish: Fall to spring 2nd gr ORF/comp: .52 3rd gr ORF/comp: .56 Spanish to English 2nd gr. Sp.ORF/EngComp: .5 3rd gr SpORF/EngBrd: .46 English to Spanish 2nd gr. EngORF/SpComp: .39 3rd gr. EngORF/SpComp: .46
	3	142						

Study	Grade	Sample size	Participants	Type of Instruction	Language Proficiency	ORF measure	Criterion Measure	Validity Evidence
Betts, J., Muyskens, P., & Marston, D. (2006). Tracking the progress of students whose first language is not English towards English proficiency: Using CBM with English language learners. *MinneTESOL/WITESOL Journal*, 23, 15–37.	2	254	L1:Spanish, Hmong, Somali L2: English	ELL	unknown	Curriculum	NALT broad reading	ORF/NALT: .69 Spanish speakers: .64 Hmong speakers: .73 Somali speakers: .69
Crosson, A. C., & Lesaux, N. K. (2010). Revisiting assumptions about the relationship of fluent reading to comprehension: Spanish speakers' text-reading fluency in English. *Reading and Writing, 23*, 475 – 494. doi:10.1007/s11145-009-9168-8	5	76	L1: Spanish L2: English	Develop. Bilingual: Spanish and English instruction	Spanish PPVT English PPVT English WLPB listening comprehension	GORT	Gates reading comprehension	GORT/GATES: .442 GORT/decoding: .62 GORT/vocab: .22
Felt, J.N. (2015). Word calling on curriculum based measures in English language learners. (Unpublished Doctoral dissertation). University of Wisconsin, Madison, Wisconsin.	3 4 5	238	L1: Spanish L2: English	Dual language immersion	ACCESS 29 level 1 78 level 2 86 level 3 27 level 4 5 level 5 13 level 6	English: DIBELS Spanish: IDEL	WKCE broad reading	DIBELS/WKCE: .63 IDEL/WKCE: .68
Grasparil, T. A., & Hernandez, D. A. (2015). Predictors of Latino English Learners' Reading Comprehension Proficiency. *Journal of Educational Research and Practice, 5*(1).	3	1376	L1: Spanish L2: English	ELL English only	CELDT listening comprehension M = 16.35 (SD=2.88) Max=20 CST vocab	Unknown	Comprehension Composite: CST, CAT6, CELDT reading	ORF/comp: .67 ORF/vocab: .61 Vocab/comp: .73

Study	Grade	Sample size	Participants	Type of Instruction	Language Proficiency	ORF measure	Criterion Measure	Validity Evidence
Jimerson, S. R., Hong, S., Stage, S., & Gerber, M. (2013). Examining oral reading fluency trajectories among English language learners and English speaking students. *Journal of New Approaches in Educational Research, 2*(1), 3-11.	Long. 1 2 3 4	85	L1: Spanish L2: English	ELL English only	LAS	ORALJ	SAT9 broad reading	1st gr to 4th gr ORF/broad: .66 2nd gr to 4th gr ORF/broad: .78 3rd gr. To 4th gr. ORF/broad: .76 4th gr. To 4th gr. ORF/broad: .72
Kim, Y. S. (2012a). The relations among L1 (Spanish) literacy skills, L2 (English) language, L2 text reading fluency, and L2 reading comprehension for Spanish-speaking ELL first grade students. *Learning and Individual Differences, 22*(6), 690-700.	1	150	L1: Spanish L2: English	ELL English only	WJ-III oral language, PPVT4, WASI vocab	DIBELS	SAT10 comprehension WRMT comprehension	ORF/SAT10: .74 ORF/comp: .76
Kim, J. S. (2012b). Evaluating the Predictive Validity of DIBELS Literacy Measures with Third Grade Spanish-Speaking English Language Learners. (Unpublished doctoral dissertation). University of California, Riverside, CA.	3 5	522	L1: Spanish L2: English	ELL English only	CELDT 34 beginner 88 early intermediate 291 intermediate 94 early advanced 15 advanced	DIBELS	CST broad reading	ORF/CST: .53 Beginner: ORF/CST: .55 Intermediate ORF/CST: .38 Advanced ORF/CST: .36

Study	Grade	Sample size	Participants	Type of Instruction	Language Proficiency	ORF measure	Criterion Measure	Validity Evidence
Millett, J. R. (2011). The Ability of Oral Fluency to Predict Reading Comprehension Among ELL Children Learning to Read (Unpublished doctoral dissertation). Arizona State University.	1 2 3	65	L1: Spanish L2: English	ELL English only	PPVT 1st gr. M=56.81 SD=16.18 2nd gr. M=68.49 SD=17.84 3rd gr. M=88.27 SD=17.80	DIBELS	TerraNova comprehension	1st gr. ORF/2nd gr. TN: .30 1st gr. ORF/2nd gr. TN: .48 2nd gr. ORF/3rd gr. TN: .66 3rd gr. ORF/3rd gr. TN: .68 3rd gr. ORF/3rd gr. PPVT: .35 3rd gr. PPVT/3rd gr/ TN: .60
Nam, J. E. (2012). An examination of the predictive validity of early literacy measures for Kroean English language learners. (Unpublished doctoral dissertation). University of California, Riverside, California.	1	102	L1: Korean L2: English	ELL English only	CELDT 3 beginning 24 early intermediate 42 intermediate 33 early advanced/ advanced	DIBELS	WRMT passage comprehension	DIBELS/comp: .82 Beginning DIBELS/comp: .87 Intermediate DIBELS/comp: .75 Advanced DIBELS/comp: .69
Pretorius, E. J., & Spaull, N. (2016). Exploring relationships between oral reading fluency and reading comprehension amongst English second language readers in South Africa. *Reading and Writing, 29*(7), 1449-1471.	1772	5	L1: various South African languages L2: English	foreign language immersion	unknown	Curriculum	Researcher created comprehension measure	ORF/Comp: .61

Study	Grade	Sample size	Participants	Type of Instruction	Language Proficiency	ORF measure	Criterion Measure	Validity Evidence
Quirk, M., & Beem, S. (2012). Examining the relations between reading fluency and reading comprehension for English language learners. *Psychology in the Schools, 49*(6), 539-553.	2 3 5	171	L1: Spanish L2: English	ELL English only	CELDT 47 beginning 89 intermediate 35 advanced	AIMSweb R CBM	Gates MacGinitie Comprehension Subtest	ORF/comp: .63
Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban, first grade students. *Reading Research Quarterly, 42*, 460 – 466. doi:10.1598/RRQ.42.4.5	1	59	L1: Spanish L2: English	ELL	Unknown	DIBELS	GRA+DE comprehension	ORF/comp: .80
Sáez, L., Park, B., Nese, J. F., Jamgochian, E., Lai, C. F., Anderson, D., ... & Tindal, G. (2010). Technical Adequacy of the easyCBM Reading Measures (Grades 3-7), 2009-2010 Version. Technical Report# 1005. Behavioral Research and Teaching.	3 4 5 6 7	91 79 88 33 148	L1: variety L2: English	unknown	Unknown	easyCBM	OAKS broad reading	3rd gr. ORF/OAKS: .51 4th gr. ORF/OAKS: .61 5th gr. ORF/OAKS: .48 6th gr. ORF/OAKS: .75 7th gr. ORF/OAKS: .51
Stokes, N. O. (2010). Examining the relationship among reading curriculum-based measures, level of language proficiency, and state accountability test scores with middle school Spanish-speaking English language learners. (Unpublished doctoral dissertation). Loyola University, Chicago, IL.	6	90	L1: variety L2: English	Unknown English only	AZELLA 2 preemergent 0 emergent 8 basic 80 intermediate 0 proficient	AIMSweb R CBM	AIMS broad reading	Fall ORF/spring AIMS: .577

Study	Grade	Sample size	Participants	Type of Instruction	Language Proficiency	ORF measure	Criterion Measure	Validity Evidence
Wiley, H. I., & Deno, S. L. (2005). Oral reading and maze measures as predictors of success for English learners on a state standards assessment. *Remedial and Special Education, 26*, 207–214.	3 5	15 14	L1: variety L2: English	Unknown	Unknown	Curriculum	MCA broad reading	3rd gr. ORF/MCA: .61 5th gr. ORF/MCA: .69
Correlational and diagnostic evidence								
Farmer, E. (2013). Examining predictive validity and rates of growth in curriculum-based measurement with English language learners in the intermediate grades. (Unpublished doctoral dissertation). Loyola University, Chicago, IL.	4 5 6	82 59 59	L1: variety L2: English	ELL	ACCESS 4th gr. Mean level 3-4, 5th grade mean level 3-4, 6th grade mean level 2-3	AIMSweb R CBM	ISAT broad reading	4th gr. Winter ORF to spring ISAT: .74 SE .96 SP .42 5th gr. Winter ORF to spring ISAT: .49 SE: .8 SP .48 6th gr. Winter ORF to spring ISAT: .68 SE .97 SP .26
Ganan, B. J. (2012). The Fluidez en La Lectura Oral (FLO) Portion of the Indicadores Dinamicos De Exito en La Lectura (IDEL) and the English Language Portion of the Illinois Standard Achievement Test (ISAT): A Correlational Study of Second and Third Grade English Language Learners. (Unpublished doctoral dissertation). Roosevelt University, Schaumburg, IL.	2 3	234	L1: Spanish L2: English	Develop. Bilingual: 70% Spanish, 30% English reading instruction by 3rd grade	unknown	Spanish: IDEL	ISAT broad reading	2nd gr. IDEL/3rd gr. ISAT: .476 SE .45 SP .87

Study	Grade	Sample size	Participants	Type of Instruction	Language Proficiency	ORF measure	Criterion Measure	Validity Evidence
Muyskens, P., Betts, J., Lau, M. Y., & Marston, D. (2009). Predictive Validity of Curriculum-Based Measures in the Reading Assessment of Students Who Are English Language Learners. *California School Psychologist, 14,* 11-21.	5	1529	L1: Spanish, Hmong, Somali L2: English	ELL	Unknown	Curriculum	MCA broad reading	Fall ORF/ spring MCA: .62 AUC: .78 SE .45 SP .9
Vanderwood, M. L., Tung, C. Y., & Checca, C. J. (2014). Predictive validity and accuracy of oral reading fluency for English learners. *Journal of Psychoeducational Assessment, 32*(3), 249-258.	2	547	L1: Spanish L2: English	ELL English only	CELDT 198 low 193 intermediate 159 advanced	DIBELS	CST broad reading	Low winter DIBELS/ spring CST: .66 SE .64 SP .90 Intermediate Winter DIBELS/ Spring CST: .52 SE .15 SP .98 Advanced Winter DIBELS/ Spring CST: .66 SE .25 SP .99
Diagnostic evidence								
Echols, J. M. Y. (2010). The Utility of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in Predicting Reading Achievement. (Unpublished doctoral dissertation). Seattle Pacific University, Seattle, WA.	1 2	228	L1: Variety L2: English	unknown	unknown	DIBELS	WASL broad reading	SE .76 SP .6

Study	Grade	Sample size	Participants	Type of Instruction	Language Proficiency	ORF measure	Criterion Measure	Validity Evidence
Hosp, J. L., Hosp, M. A., & Dole, J. K. (2011). Potential bias in predictive validity of universal screening measures across disaggregation subgroups. *School Psychology Review, 40*(1), 108.	1 2 3	403 311 247	L1: Variety L2: English	unknown	unknown	DIBELS	UCRT broad reading	1st gr. AUC .86 SE .74 SP .82 2nd gr. AUC .84 SE .84 SP .60 3rd gr. AUC .85 SE .91 SP .47
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.	Long. 1-3	901	L1: Variety L2: English	Unknown	PPVT	DIBELS	SAT10 comprehension	1st gr. To 3rd gr. SE .9 SP .45