The Pennsylvania State University

The Graduate School

INTERRATER RELIABILITY OF CURRICULUM-BASED MEASURES IN READING (R-CBMS) FOR ENGLISH LEARNERS

A Dissertation in

School Psychology

by

Ashley Marinez

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

May 2020

The dissertation of Ashley Marinez was reviewed and approved* by the following:

Barbara A. Schaefer

Associate Professor of Education, School Psychology

Dissertation Adviser

Chair of Committee

Shirley A. Woika

Professor of Education, School Psychology

Katie E. Hoffman

Teaching Professor of Education, Counselor Education

Janet Van Hell

Professor of Psychology and Linguistics

Co-Director of the Center for Language Science

James C. Diperna

Professor of Education, School Psychology

Professor in Charge of School Psychology

*Signatures are on file in the Graduate School

ABSTRACT

The purpose of the current study was to determine interrater reliability (IRR) of Oral Reading Fluency (ORF) Curriculum-based Measures (R-CBM) when used with Spanish-speaking English Learner (EL) students. The ORF R-CBM probes obtained from AIMSweb are measures of a student's reading accuracy skills and reading fluency skills. Certified school psychologists who have graduated from The Pennsylvania State University School Psychology program, current school psychologists in the field, and current school psychology students and faculty from various other graduate training programs were recruited as raters. Participants were asked to score several ORF passages twice. During the first round of scoring (T1), raters followed the standard directions provided by the AIMSweb program. About two months later (T2), raters scored the four additional passages using more comprehensive instructions that targeted issues presented by EL students. The sample demographics and the distribution of ratings for the ORF passage for both administrations were analyzed (e.g., *M*, *SD*, range). IRR was calculated for T1 and T2 using the two-way, random-effects intraclass correlation (ICC) approach. Lastly, the average difference and consistency for passages between T1 and T2 were tested for significance using a two-way, mixed-effects model. A total of 35 raters scored 8 EL student passages over both administrations, resulting in 70 ratings. Overall, good to excellent interrater reliability emerged when scoring the ORF R-CBMs for both T1 and T2, indicating statistically significant IRR indices ($p < .05$). Results provide evidence to support the use of R-CBMs with Spanish-speaking EL students.

Keywords: EL, R-CBM, interrater reliability

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

Problem Statement

Introduction

English Language Learners (ELLs) or English Learners (ELs) are common terms used to define students who are learning English (Gersten et al., 2007; Klingner, Hoover, & Baca, 2008). In the U.S., approximately 4.6 million ELs are students currently enrolled in public school (McFarland et al., 2017). Language proficiency can vary across students and the language barrier presented can lead EL students to have different academic needs compared to native English speakers (Klingner et al., 2008; Rhodes, Ochoa, & Ortiz, 2005). Although linguistically and culturally diverse assessments are ideal when identifying an EL student's needs, such measures can be difficult to obtain and difficult to administer (Salvia, Ysseldyke, & Bolt, 2013)

One possible solution for assessing an EL student would be to use Curriculum-based Assessments (CBAs). CBAs measure a student's current skill level by taking relevant test items directly from instructional materials to ensure the student understands the task (Rhodes et al., 2005). Various forms of CBAs are available that measure both broad and narrow skill sets (Klingner et al., 2008; Salvia et al., 2013; Shapiro, 2011). According to Klingner et al. (2008), the most frequently used CBA is known as a Curriculum-based Measure (CBM). Similar to CBAs, CBMs provide a quick, simple, and inexpensive way of collecting student data for most academic areas (Codding, Petscher, & Truckenmiller, 2015; Deno, 1985; Deno, 2003; Fuchs, 2016; Shapiro, 2011). Unlike most CBAs, however, CBMs provide normative data with strong reliability and validity to support its use in schools (Baker & Good, 1995; Codding et al., 2015; Deno, 1985; Deno, 2003; Espin & Wallace, 2004; Fuchs, 2016; January & Ardoin, 2015; Salvia et al., 2013; Shapiro, 2011).

CBMs are often used as universal screeners and as progress monitoring tools to help identify and assist students with skill deficits (Deno, 2003; Fuchs, 2016; Salvia et al., 2013; Shapiro, 2011). Several studies have strongly supported the use of Reading-CBMs (R-CBMs) in particular, showing that these reading-based measures reliably predict later achievement (Chiappe, Siegel, & Wade-Woolley, 2002; Gersten et al., 2007; Geva, Yaghoub-Zadeh, & Schuster, 2000; Lafrance & Gottardo, 2005; Swanson, Sáez, & Gerber, 2004). R-CBMs are sensitive enough to measure student growth and to differentiate between language barriers and learning disabilities for EL students (Deno, 2003; Espin & Wallace, 2004; Espin, Wallace, Lembke, Campbell, & Long, 2010; Graves, Plasencia-Peinado, Deno, & Johnson, 2005; January & Ardoin, 2015; Miller, Bell, & McCallum, 2015; Nese, Park, Alonzo, & Tindal, 2011). Many studies have also demonstrated that R-CBMs can reliably predict EL student achievement on high-stakes assessments (Codding et al., 2015; Kim, Vanderwood, & Lee, 2016; Ramírez, Domínguez, & Shapiro, 2007).

While CBMs can be extremely useful, educators should be aware of a number of considerations before implementing such measures. This includes the variability of materials across teachers, the reliance on visual data, the dangers of teaching-to-the-test, representativeness of the normative sample, and the lack of variability in methods between studies (Deno, 2003; Rhodes et al., 2005; Shapiro, 2011). One issue in particular that may affect EL students' performance is the lack of scoring instructions provided by CBM developers for students with accents. On most R-CBMs, for example, mispronunciations are to be counted as errors whereas dialectical differences are not (Shinn & Shinn, 2002; Wright, 2013). While this directive may seem clear, this does not specify whether mispronunciations resulting from accents should be scored as errors. As such, interrater reliability (IRR) is brought into question when looking at EL

student performance on R-CBMs. To date, little research has examined the degree of IRR of R-CBMs when given to EL students.

Rationale and Research Questions

The current study aimed to examine IRR of Oral Reading Fluency (ORF) R-CBM passages when used with an EL student. The focus of this study was to observe how school personnel rate student oral reading performance and to establish purposed "gold-standards" for reliably scoring ORF R-CBMs when a student's accent may complicate scoring. The primary research questions were as follows:

1. To what degree do raters differ from one another when scoring ORF R-CBM probes read by Spanish-speaking EL students?
2. When raters are given improved, highly detailed and specific guidelines, does IRR increase?

Due to the lack of clear scoring instructions that specifically relate to EL students, it was anticipated that raters would initially differ when scoring EL students' ORF R-CBMs, resulting in lower IRR. It was also believed that IRR would increase significantly when clearer guidelines for scoring Spanish ORF R-CBMs are provided.

CHAPTER 2

Literature Review

English Learners (ELs)

Klingner et al. (2008) define an English Language Learner (ELL) as a student who has a native language other than English and who is "in the process of acquiring English as a second language" (p. 5). In 2015, the term ELL changed to English Learner (EL) with the signing of the Every Student Succeeds Act (ESSA) so as to better incorporate all students who are learning English regardless of their proficiency levels and previous languages learned (Gersten et al., 2007, Public Law 114-95). According to Rhodes et al. (2005), Title VII of the Improving America's School Act of 1994 (Public Law 103-382) defines a student with Limited English Proficiency (LEP) as someone who is having significant difficulties learning English and was born outside the U.S., has a native language other than English, is Native American or Alaskan Native with a native language other than English, or is migratory and has a native language other than English. This includes "speaking, reading, writing, or understanding the English language" (Public Law 103-382).

In 1993, over 2,430,000 LEP students were enrolled in U.S. public schools, resulting in an 8.8% increase since 1986 (Henderson, Abbot, & Strang, 1993). The percentage of EL students attending public schools in the U.S. has since increased from 9.1% (i.e., about 4.3 million) during the 2004-2005 school year to 9.4% (i.e., about 4.6 million students) in the 2014-2015 school year (McFarland et al., 2017). For the 2014-2015 school year, approximately 10% of the student body consisted of EL students in seven states. Of all the EL students in the U.S., Spanish-speaking students have been the most predominant (Klingner et al., 2008; Salvia et al., 2013; Sandberg & Reschly, 2011). In 2000, 77% of EL students consisted of Spanish speakers

(Rhodes et al., 2005). According to the most recent census from school year 2014-2015, 3.7 million are Spanish-speaking EL students, representing about 77.1% of all EL students and about 7.6% of all students in U.S. public schools (McFarland et al., 2017). In comparison, the second most common language for EL students around the U.S. is Arabic, making up only 2.3% of all EL students (McFarland et al., 2017). Given the number of EL students in U.S. public schools, special attention should be given to assisting students who are having difficulties learning English.

The No Child Left Behind Act (NCLB, 2002) requires that all schools monitor student progress and that all students make adequate yearly progress (Nese et al., 2011; Sandberg & Reschly, 2011). This includes EL students, ethnic minority students, students from low socioeconomic backgrounds, and students receiving special education (Nese et al., 2011). During the 2014-2015 school year, about 13.8% (about 665,000 EL students) were also identified as students with disabilities (McFarland et al., 2017). As compared to their White counterparts, EL students have higher rates of underachievement in schools and greater rates of representation in special education (Klingner et al., 2008; Spinelli, 2008). The difference in achievement between native English speakers and EL students may be explained in part by the difference in educational support and academic needs that these students require.

Critical Academic Components for EL Students

Cummins (1999) suggests a clear distinction between language acquisition and development between conversational and academic language. He explains that an EL student's conversational language, i.e., Basic Interpersonal Communicative Skills (BICS), usually develops before academic language, or Cognitive Academic Language Proficiency (CALP). While the focus of teaching BICS before CALP is clearly seen in the structured play of daycare

centers, preschools, and kindergarten classrooms, BICS is less emphasized after the first grade. Laija-Rodríguez, Ochoa, and Parker, 2006 (2006) note that CALP is often described as a milestone to English language development in ELs and "as having a significant relationship with academic achievement" (p. 87). When instructing EL students, teachers often focus on developing the student's CALP and place little emphasis on the student's BICS (Collier, 1998; Laija-Rodríguez et al., 2006). According to Cummins (1999), this emphasis may actually make it more difficult for an EL student to develop English language proficiency; EL students who do not have time to develop their BICS can find it more difficult to catch up to their peers.

To help further support the development of English language for ELs, Cummins (1999) states that educators should be focusing on teaching both BICS and CALP and not choosing one over the other. In addition, Cummins (1999) adds the CALP development should address three major constructs: cognition, academics, and language. Cognition refers to providing cognitively challenging activities that require higher-order thinking skills rather then memorization and rote application skills. Academics refers to providing content-based instruction in programs such as English as a Second Language (ESL). The final construct, language, refers to opportunities for comparing the students' native language to "their community's language use, practices, and assumptions" (Cummins, 1999, p. 6). By incorporating the above components, BICS and CALP, into instruction, teachers can use the students' native language to support their English language development.

When working with an EL student, the particular needs of the student likely will vary from those of a native English speaker (Collier, 1998). Klingner et al. (2008) discuss five major areas of reading that can impact an EL's English acquisition. These areas are known as the "big five" of reading (Vanderwood, Tung, & Hickey, 2014, p. 181); phonological awareness,

alphabetic principle, fluency, vocabulary, and reading comprehension. Phonological awareness refers to the skills needed to both distinguish and manipulate spoken language. For EL students, phonemes in their first language may overlap very little or be dissimilar to those presented in the second language, which can result in more difficulties with English acquisition. The second area of reading, alphabetic principle, refers to one's ability to identify letter names and their corresponding letter sounds. While some languages do use the same letters, languages like Chinese, Greek, Russian, and Arabic have very different written characters that do not correspond well with the English language. Some languages also have additional letters and sounds that are added into the alphabet. The Spanish alphabet, for example, has letters that are not in the English alphabet (e.g., ñ). Even when using the same alphabet, different languages have distinct names and sounds for each letter, making it more difficult for EL students to separate English letter names and sounds from those in their native language (Klingner et al., 2008, p. 61).

Fluency is the third area discussed by Klingner et al. (2008). Fluency is a student's "ability to read quickly and accurately, with expression" (Klingner et al., 2008, p. 63). EL students tend to struggle with fluency due to difficulties with other areas of reading (e.g., phonological awareness, reading comprehension, etc.). Even when EL students become proficient in English, their accents may influence their fluency, leading teachers to believe that the student has a reading fluency deficiency (Derwing & Munro, 2005). Derwing and Munro (2005) looked at the importance of mutual intelligibility of language between students and their teachers. Mutual intelligibility refers to how well a teacher and an EL student are able to understand each other. Derwing and Munro (2005) note that a large portion of ESL teachers (i.e., 67%) are underprepared to teach EL students proper English pronunciations due to the

phonological differences between English and the student's native language. This may be due to a lack of understanding of phonetics in other languages. As a result, mutual intelligibility between the EL student and the teacher is low and EL instruction is not individualized to meet the student's needs. Overall, Derwing and Munro (2005) suggest seeking opportunities for professional development that focuses on teaching English pronunciation by using the student's native language as bases for support.

The area of vocabulary refers to one's ability to understand the individual words in a language and understand how sentences are formed (Klingner et al., 2008). When EL students are able to pronounce and define a word correctly, this does not always indicate a strong understanding of how it should be used in a sentence. Sentence composition can differ greatly from the students' native language, adding another barrier to their ability to transfer their prior vocabulary skills from their native language to the new language. For example, the phrase "the black cat slept" in English becomes "el gato negro durmió" in Spanish. When translated back to English, the Spanish phrase has changed to "the cat black slept". These minor translation alterations can confuse a student who lacks the proper skills needed to switch between the two languages effectively (Ordóñez, Carlo, Snow, & McLaughlin, 2002). Ordóñez et al. (2002) looked at how well students were able to provide accurate word definitions, object descriptions, and sentence composition in both English and Spanish. From their study, Ordóñez and colleagues concluded that such skills only transferred between languages when the students' vocabulary knowledge and communication skills were at equivalent levels for both English and Spanish. Overall, strengthening EL students' vocabulary skills is crucial in helping them understand sentence formation in the English language and succeed academically.

Finally, reading comprehension refers to one's ability to understand what he/she is reading (Klingner et al., 2008). Alyousef (2006) adds, "reading comprehension is a combination of identification and interpretation skills" (p. 63). The prior four reading components discussed all affect EL students' ability to understand what they have read. Their ability to identify letter names and sounds, their ability to pronounce and understand words correctly, and their ability to combine those words fluently affect how they understand the passage. Without the other components, reading comprehension can be very difficult to impossible (Klingner et al., 2008). It is important to note, however, that one should not automatically assume that reading comprehension skills are proficient even when most of the other skills are strong. In a study looking at the relationship between reading fluency and reading comprehension, Quirk and Beem (2012) found that 55% of EL students who performed in the average to above average range on one task performed significantly lower on the other. More specifically, EL students tended to score significantly higher on the reading fluency task than on the reading comprehension task. Quirk and Beem (2012) comment that while reading fluency and reading comprehension are significantly related, educators should be cautious when determining skill adequacy in EL students.

Rhodes et al. (2005) note two additional components to be aware of when identifying an EL student who may need additional services. The first is variations in proficiency level for both languages. Due to variability in student background and experiences, students may have substantially variable degrees of proficiency in both their native language and in English (Collier, 1998). For example, while student's have lived in Puerto Rico their whole life and have received comparable instruction in both languages, their English proficiency may be limited

while their Spanish proficiency might be advanced given that in their environment the predominant language of their friends and family is Spanish.

The second component to remember is a student's bilingualism (Collier, 1998; Rhodes et al., 2005). This refers to how both languages were learned (i.e., sequential versus simultaneous), as well as why the languages were learned (i.e., elective versus circumstantial). If a student learns one language first and then subsequently learns the second language, the bilingual pattern would be sequential. If a student learns a new language by choice, this would be considered elective bilingualism. On the contrary, students who learn both languages at the same time because their family speaks both languages equally would be considered a simultaneous and circumstantially bilingual. Learning these possible patterns of language acquisition may help identify the student's individual needs. Being aware of these key components when formally assessing EL students may help determine the degree of academic assistance required for the child to succeed.

Goldenberg (2013) detailed common, beneficial instructional practices and recommendations used when working with EL students. From these, four major principles were identified. The first is that generally effective practices used with native English speakers are likely to be effective with ELs as well. Such practices include providing clear goals and objectives, ensuring materials are appropriate and challenging, creating well-designed instruction and instructional routines, providing informative feedback, and effectively modeling skills, strategies, and procedures in the classroom. In addition to using common instructional practices, the second major principle stated that ELs require additional instructional supports (Goldenberg, 2013). Some instructional strategies proposed included the use of pictures, demonstrations, and real-life objects to provide hands-on experience. Additionally, interactive learning activities,

scaffolding, repetition of information (i.e., gestures, visual cues), pre- and post-assessments, introducing new vocabulary in an integrated manner, and providing differentiated instruction according to the student's English proficiency level were found to be effective (Goldenberg, 2013; Hansen, 2006; Slavit & Ernst-Slavit, 2007; Walqui, 2006).

The third principle presented by Goldenberg (2013) encouraged the use of the EL's native language to promote academic development. This principle focuses on teaching academic content and skills in the student's native language first to later support English instruction and activities. For example, teachers can provide definitions and brief explanations of content in the student's native language prior to continuing instruction in English. The final major principle noted is that ELs need early and ample opportunities to develop proficiency in English (Goldenberg, 2013). This includes repeated practice, the use of scaffolding to support the additional practice, and pre- and post-assessment (Goldenberg, 2013; Hansen, 2006; Walqui, 2006). While various programs have been developed to support the proposed principles above, the strategies listed can be used on their own to support the learning needs of EL students (Goldenberg, 2013; Hansen, 2006; Turuk, 2008; Walqui, 2006).

Walqui (2006) and Spinelli (2008) both suggest that the best way to support EL students' language development in schools is to model teaching strategies according to Vygotsky's social-cultural theory. According to his theory, Vygotsky argues that social interactions are the most crucial aspects of learning and development (Turuk, 2008; Walqui, 2006). Scaffolding and mediation are seen as the primary strategies used to learn new information and the students' zone of proximal development helps determine what the students do not know yet, what they are currently learning, and what they have mastered (Spinelli, 2008; Turuk, 2008; Walqui, 2006). As such, teachers can focus their support on the EL student' s zone of proximal development by

implementing scaffolding and mediation opportunities (Hansen, 2006; Spinelli, 2008; Turuk, 2008; Walqui, 2006).

Vanderwood et al. (2014) suggest that student academic achievement may actually rely heavily on the American culture itself and on the student's English proficiency levels. In addition, both Sandberg and Reschly (2011) and Vanderwood et al. (2014) briefly discuss some additional obstacles that culturally and linguistically diverse students often face, including socio-economic disadvantages underrepresentation in the normative samples for most standardized assessments, and a lack of English academic language (i.e., CALP). As such, it is important to consider the student at an individual level to understand his or her particular needs. Besides the supports listed above, Vanderwood et al. (2014) add that individualized, explicit instruction is a major key to helping EL students succeed. Furthermore, English proficiency assessments are critical in the decision-making process to understand how the student is progressing with the supports he or she have in place.

Linguistically and Culturally Diverse Measures

Stokes (2010) believes that EL students' needs are not effectively being met by the current educational system. She notes "there is a need for a system in which students at risk of reading failure are identified early in the school year and are then provided with appropriate intervention in order to remediate those weaknesses and improve reading achievement" (Stokes, 2010, p. 120). As such, many districts have begun to incorporate multi-tiered systems of support, like Response to Intervention and Instruction (RTII), in order to provide early services (Kondisko, 2017; Salvia et al., 2013; Stevenson, 2015; Stokes, 2010). In the RTII model, student needs are addressed with evidence-based interventions through a multi-tiered system of supports prior to a formal assessment process. In order to implement programs like RTII with fidelity,

continuous progress monitoring is required (Fuchs, 2016; Salvia et al., 2013; Shapiro, 2011). Even so, RTII implementation alone is not enough if evaluation and assessment measures are not reliably and validly distinguishing which EL students need more assistance.

Linguistically and culturally diverse measures would be ideal in determining what an EL student needs the most from instruction; however, several obstacles prevent such measures from being readily available (Vanderwood et al., 2014; Wayman, McMaster, Sáenz, & Watson, 2010; Spinelli, 2008). Salvia et al. (2013) note three main issues related to developing culturally and linguistically sensitive assessments. The first is the number of different languages spoken and the low prevalence rates of some languages. Nationally, over 400 different languages are currently spoken in the U.S. (Rhodes et al., 2005). For the 2016-2017 school year, approximately 51.48% of students spoke a language other than English in New York City's public school system alone, resulting in over 155 different languages spoken in all five boroughs (New York City Department of Education, 2017; Sandberg & Reschly, 2011). Given the numerous languages and dialects spoken, the costs associated with creating assessments in each language, and the low prevalence of many of those who speak certain languages, Salvia et al. (2013) argue that creating assessments for each language is cost prohibitive and establishing sound psychometric norms would be difficult (Salvia et al., 2013).

The second issue presented is the limited availability of school personnel that would be able to competently and fluently give the assessments in the student's native language (Salvia et al., 2013). Even if the measures were available in the student's native language, often few educators can fluently speak the student's native language (Spinelli, 2008). Lastly, dialect variability exists across students even when the same language is spoken. Prevalent languages like Spanish can differ greatly depending on the country and region of origin of the speaker,

leading to vastly different social influences, cultural backgrounds, and dialects. These variations can make it difficult to assess students in a different language, even if they indicate that they speak the same language.

Curriculum-Based Assessments

Vanderwood et al. (2014) suggest that one possible solution for establishing useful methods for assessing EL students' academic needs is to use a more direct measure of classroom performance, such as Curriculum-based Assessments (CBAs). Rhodes et al. (2005) define a CBA as a measure of a student's instructional need through the use of classroom curricula. The goal of a CBA is to understand where the student's current skills lie (i.e., frustration, instructional, or mastery level) in a specific academic area, such as reading and mathematics (Hosp, Hosp, & Howell, 2016). Students are assessed using the same classroom materials with which they are familiar, allowing the teacher to directly monitor student progress while ensuring that the students understands what is presented to them. CBAs also allow for more flexibility for the teacher to adjust materials and create more linguistically and culturally appropriate and relevant assessment tools (Hosp et al., 2016; Klingner et al., 2008; Sandberg & Reschly, 2011).

Types of Curriculum-Based Assessments

Various CBA measures have been created to provide a wide variety of tools with different academic foci. The broadest measure is known as a General Outcome Measure (GOM). GOMs are tools used to assess academic outcomes that require multiple skills to be used (Hosp et al., 2016; Shapiro, Keller, Lutz, Santoro, & Hintze, 2006; Vanderwood et al, 2014). An example of a common GOM is a test for oral reading skills. The oral reading skills measure would require the student to use multiple reading skills learned, including identification of letter sounds, blending of letter sounds, correctly producing words, fluently reading sentences, and

combining the meaning of the content read (Salvia et al., 2013). GOMs are standardized measures, providing acceptable reliability and validity levels (Shapiro, 2011). Salvia et al. (2013) note that GOMs are often used as a way to measure student progress towards long-term goals (p. 109).

Similar to the GOM, Skill-Based Measures (SBMs) are also used to measure long-term goals (Hosp et al., 2016). The difference between SBMs and GOMs lies in the amount of skills assessed with one task. While GOMs require the coordination of many subskills, SBMs require fewer subskills. This allows for more isolation of each subskill, providing a more direct assessment of each individual skill learned (Salvia et al., 2013). An example of an SBM is a single-digit math computation probe, where the student's ability to add and subtract single-digit numbers is assessed. Two similar tools are the Mastery Measures (MMs) and the Subskill Mastery Measures (SMMs). MMs are diagnostic evaluation tools that look at how well a student has mastered a specific skill while SMMs usually focus on screening students in a particular academic area, measuring progress of that particular skill towards a short-term goal (Hosp et al., 2016; Salvia et al., 2013). An example of an SMM is a single-digit, addition math task, where the student's ability to add single-digit numbers is measured. Using this same task, an MM tool would only assess whether the student has mastered single-digit addition. However, it should be noted that MMs and SMMs are not usually standardized measures of student performance and should be interpreted with caution (Codding et al., 2015; Hosp et al., 2016; Kim et al., 2016; Ramírez et al., 2007; Shapiro, 2011).

As mentioned above, one issue with using the CBAs listed is the lack of standardization (Spinelli, 2008). While a quick understanding of a student's current academic level is obtained, comparison to peers and normative data is rarely available. Even though GOMs may provide

standardized means to assess academic performance, those measures are broad and assess many different academic areas at once (Hosp et al., 2016; Salvia et al., 2013). This makes it difficult for the teacher to know what specific skill deficits the student may have. To provide a more standardized measure of a student's academic performance in a particular area, Curriculum-based Measures (e.g., CBMs) are more frequently used (Hosp et al., 2016; Klingner et al., 2008; Spinelli, 2008). CBAs are not usually standardized and are often drawn from a teacher's personal classroom materials while CBMs have been normed and tested for reliable scores and valid interpretations (Klingner et al., 2008). A CBM is a form of CBA that provides more valid and reliable results, are often norm referenced, and provide a greater understanding of how the student is performing compared to others (Shapiro et al., 2006). CBMs also provide instructions for both standard administration and scoring procedures. CBMs are a simple, brief, time efficient, reliable, valid, easily understood, and an inexpensive way of obtaining student achievement information in a wide-variety of skills (Codding et al., 2015; Deno, 1985; Deno, 2003; Fuchs, 2016; Hosp et al., 2016; Shapiro, 2011; Spinelli, 2008; Vanderwood et al., 2014).

Curriculum-Based Measures

CBMs can be an invaluable tool in determining the instructional needs of an EL student (Vanderwood et al., 2014). CBMs provide many advantages for educators as well (Deno, 1985, Deno, 2003, Fuchs, 2016, Hosp et al., 2016; Mennuti, Christner, & Freeman, 2012; Rhodes et al., 2005; Salvia et al., 2013). Some advantages include improved communication of results due to its simple and clear data collection, increased sensitivity to particular academic skills, improved normative data that includes low-incident populations, and cost effectiveness (Deno, 1985). According to Rhodes et al. (2005), some other advantages of using CBMs in schools include the use of instructional content for a more direct assessment of skills, the ability to individualize

materials for student needs (i.e., cultural, educational, and linguistic needs), and the capability to progress monitor a student's performance. Deno (1985) notes that CBMs can be curriculum-referenced (i.e., direct measure of student performance on local school curricula data), individual-referenced (i.e., student is compared to himself or herself for progress monitoring), and peer-referenced (i.e., student performance can be compared to his or her peers). This flexibility increases the utility of CBM probes and allows for multiple methods of analyzing student progress.

In addition, CBMs provide teachers with a quick and easy method for "assess[ing] the effects of their instruction on individual students" (Fuchs, 2016, p. 5). From the data collected, teachers can then make data-based instructional decisions for students depending on their individual needs (Fuchs, 2016). Because of this, CBMs also fit well into a multi-tiered system, such as the RTII system (Kondisko, 2017; Salvia et al., 2013; Stevenson, 2015; Stokes, 2010; Vanderwood et al., 2014). CBMs provide a quick and simple method of collecting and assessing student data within the RTII model to ensure that student progress is being made (Kondisko, 2017; Vanderwood et al., 2014).

Technical Adequacy of Curriculum-Based Measures

While CBMs were initially designed to assess reading and spelling skills, CBMs now broadly assess many more academic areas through a variety of tasks, including reading, writing, spelling, and mathematics curriculum used in classrooms with specific administration and scoring guidelines (Salvia et al., 2013; Shapiro et al., 2006; Vanderwood et al., 2014). Regardless of the subject, CBMs use instructional materials to provide technically adequate and standardized measures of student academic performance (Codding et al., 2015; Deno, 1985; Deno, 2003; Fuchs, 2016; Salvia et al., 2013; Vanderwood et al., 2014). A substantial amount of

research has supported the use of CBMs (Shapiro, 2011; Shapiro et al., 2006). According to Fuchs (2016), Espin and Wallace (2004) collected information on 585 studies and reported on the utility and technical adequacy of CBMs. Of the 585 reports, 141 of the published pieces "reported empirical studies addressing questions of technical adequacy, the logistics of implementation, and instructional utility in reading, writing, spelling, and mathematics" (Fuchs, 2016, p. 6). Such evidence demonstrates the extensive use of CBMs and the empirical support for the use of such measures.

These brief measures are not only easy to learn, but they also typically provide reliable and valid scores (Fuchs, 2016; Hosp et al., 2016; January & Ardoin, 2015; Salvia et al., 2013; Shapiro, 2011; Vanderwood et al., 2014). More specifically, Shapiro (2011) notes that CBMs present with strong test-retest, internal, and interrater reliability (IRR), as well as concurrent validity, "with standardized, norm-referenced tests in reading, spelling, and written language" (p. 46). Furthermore, Shapiro (2011) notes that while mathematics CBMs have good validity and can be useful, they are not as technically sound as reading and writing CBMs. Espin, McMaster, and Rose (2012) also note that "CBMs are sensitive to growth" (p. 189), making it simple for teachers to assess if a student is making progress in a particular area or with a particular skillset.

Utility of Curriculum-Based Measures

Fuchs (2016) discusses the basics for using CBMs, the research behind CBMs, the influence that CBMs have had in education over the last three decades, and the issues raised for those with learning disabilities. Through the use of standardized administration and scoring procedures, CBMs provide a quick indicator of students who may be academically at-risk, thus they are often called an "academic thermometer" (Fuchs, 2016; Salvia et al., 2013, p. 108). Various forms of the same measure are also available, allowing for repeated sampling of a

student's performance over time (Deno, 2003; Fuchs, 2016; Hosp et al., 2016; Shapiro, 2011). CBMs are most often used as universal screeners and to monitor student progress throughout the entire school year by comparing student performance to peers on grade-level materials (Fuchs, 2016; Hosp et al., 2016; Salvia et al., 2013; Sandberg & Reschly, 2011; Shapiro, 2011). One major reason for obtaining data from CBMs is to progress monitor (Hosp et al., 2016; Fuchs, 2016). Progress monitoring helps determine the effectiveness of instruction or the effectiveness of an intervention in improving a student's academic performance (Shapiro, 2011).

In a practice guide produced by the What Works Clearinghouse (WWC) focusing on evidence-based literacy and instruction for EL students, Gersten et al. (2007) describe 21 empirical studies that have shown strong support for the use of Reading CBMs (R-CBMs) as reliable progress monitoring tools for EL students. From these studies, Gersten et al. (2007) selected particular skill sets measured by CBMs that have shown the greatest affect on EL student academic achievement. For example, R-CBMs in kindergarten and first grade have reliably shown that the development of phonological awareness skills is similar for native English speakers and EL students, and such skills can reliably predict later reading achievement (Chiappe et al., 2002; Geva et al., 2000; Lafrance & Gottardo, 2005; Swanson et al., 2004). The Oral Reading Fluency (ORF) R-CBMs in particular have been shown to be valid screening and progress monitoring measures for EL students between second and fifth grade (Gersten et al., 2007). EL student performance on the ORF task has reliably helped identify students who have a learning disability and has accurately predicted their performance on comprehensive and standardized reading assessments (Baker & Good; 1995; Domínguez de Ramírez & Shapiro, 2006; Gersten et al., 2007; Wiley & Deno, 2005).

Shapiro (2011) adds that progress monitoring is most effective when short- and long-term academic goals are used. As such, CBM data is often used to inform educational decisions (Deno, 2003; Hosp et al., 2016; Miller et al., 2015; Nese et al., 2011; Vanderwood et al., 2014). Deno (2003) discusses some educational decisions that may rely on CBM data. These decisions include "screening to identify, evaluating prereferral interventions, determining eligibility for and placement in remedial and special education programs, formatively evaluating instruction, and evaluating reintegration and inclusion of students in mainstream programs" (Deno, 2003, p. 184). Deno (2003) and Hosp et al. (2016) also note that CBMs can be used as an outcome measure.

One evaluation method that relies heavily on data obtained from CBMs is a Curriculum-Based Evaluation (CBE). A CBE is an educational evaluation that uses "task analysis, skill probes, direct observation, and other evaluation tools" to help determine a student's eligibility for specialized academic intervention (Shapiro, 2011, p. 19). While a CBE can include a Psychoeducation Evaluation (i.e., an evaluation with the purpose of determining eligibility for special education programming that will warrant an IEP), a CBE most often refers to a smaller-scale evaluation for understanding if a particular intervention is working for a student (Vanderwood et al., 2014). According to Shapiro (2011), a CBE involves five steps. The first is identification, where relevant background information, performance, and difficulties of the student are defined in order to create a hypothesis explaining possible reasons for why the student is struggling. The second is assuming causal development, or taking action on the hypothesis by providing services according to the initial findings. The third step is to assess the effects of the intervention(s) being used to either validate or modify the hypothesis. In this step, a student's academic progress is monitored to determine whether or not the intervention is

working. Making a summative decision (i.e., modifying or validating hypotheses) and determining the student needs according to the results in step three are the fourth and fifth steps of a CBE. The CBE process ensures that decisions are being driven by student data, thus, are informed and data based (Shapiro, 2011).

Predictive Abilities of Curriculum-Based Measures

Espin et al. (2010) conducted a study where secondary-school students were given both the ORF and the Maze R-CBMs in order to assess the measures' abilities to predict performance on the reading portion of a required high-stakes test; the Minnesota Basic Standards Test (MBST, Minnesota Department of Education, 2018). The ORF is a 1-minute probe that requires a student to read a passage aloud. The rater then calculates the total number of words read correctly and divides it by the total number of words attempted to obtain the reading accuracy percentage. The Maze R-CBM task requires students to read a passage and fill in the missing words (Pearson, 2012). For each missing word, the student selects from one of three words presented that best completes a sentence.

A total of 236 eighth-grade students from two middle schools took part in the CBM portion of the study in the fall, 16 percent of which were EL students who did not receive language support services. The MBST was then administered to students in February to assess the relationship between both measures. Overall, Espin et al. (2010) found that both R-CBMs were "reliable and valid predictors of performance on the state standards tests, with validity coefficients above .70" (p. 60). While strong predictive ability for the CBMs was found for the student body as a whole, it should be noted that EL student performance was not specifically analyzed to understand if the predictive ability of their CBMs differed when compared to their English-speaking counterparts.

January and Ardoin (2015) looked at the acceptability and concurrent validity of using an ORF R-CBM and the Measures of Academic Progress (MAP) assessments as academic screeners. While looking at students in the first through fifth grade, January and Ardoin (2015) found concurrent validity between ORF and MAP measures. January and Ardoin (2015) state that R-CBMs measuring ORF are not only good predictors of reading comprehension, but they can also help students make greater gains in reading when used to monitor student progress. In a similar study, Miller et al. (2015) looked at the predictability of an R-CBM, known as the Monitoring Instructional Responsiveness: Reading (MIR:R), on student performance for the Tennessee Comprehensive Assessment Program (TCAP) Reading Composite. The MIR:R measured both reading rate (i.e., total words read in 3 minutes) and reading comprehension skills in students. The study included 448 third-grade students from eight elementary schools. Overall, the study found that MIR:R comprehension rates provided strong sensitivity (85%) and specificity (53%) in predicting TCAP reading composite scores (Miller et al., 2015). Results also indicated that, while MIR:R comprehension scores had moderate predictive validity, students' reading rate was less predictive (Miller et al., 2015).

In a study conducted by Nese et al., (2011), the relationship between EasyCBM reading benchmark tools and statewide assessments of 13 states in the U.S. were assessed. Benchmarking CBMs are often given three times a year (i.e., fall, winter, and spring) and are then analyzed to see if the students are meeting academic performance standards (i.e., making adequate yearly progress). About 1,800 fourth- and fifth-grade students were assessed during the 2009-2010 school year using several R-CBM benchmarks, including Passage Reading Fluency (PRF), Vocabulary, and Multiple Choice Reading Comprehension (MCRC). Similarly to ORF tasks, PRF probes assess a student's oral reading accuracy and fluency skills (EasyCBM, 2016).

Vocabulary R-CBMs assess a student's knowledge of words and oral language skills and MCRC probes measure a student's reading comprehension skills (EasyCBM, 2016). The students' scores on the CBMs were then compared to their performance on the Oregon Statewide Achievement assessment, known as the Oregon Assessment of Knowledge and Skills (OAKS), from the previous year. According to Nese et al. (2011), the study showed "strong concurrent validity" between the CBM measures and performance on the statewide assessment (p. 608). Moreover, Nese et al. (2011) found that the vocabulary CBM measure was the strongest predictor of academic performance for both grades.

Utility of Curriculum-Based Measures With EL Students

R-CBMs have reliably been used to identify word recognition and nonsense word identification difficulties in first graders, distinguish students with learning difficulties, to predict student performance in high-stakes assessments, and to assess a student's degree of growth in specific content areas (Deno, 2003; Espin et al., 2010; Espin & Wallace, 2004; Healy, 2007; January & Ardoin, 2015; Miller et al., 2015; Moore, 1997; Nese et al., 2011; Sandberg & Reschly, 2011). Kondisko (2017) studied writing CBMs to see if they were valid for racially diverse students. More specifically, the relationship between CBM writing measures and racial categories was examined. For the study, students in grades one through three were given several writing CBM probes. The mean difference between each racial group was then compared. For this study, the racial groups consisted of Caucasian, African American, Hispanic, Asian, and multiracial students. Preliminary analyses revealed that "65% of the racial comparisons made were insignificant", suggesting that the writing CBM measures were not racially biased (Kondisko, 2017, p. 110).

To further examine the validity and reliability of CBM use with EL students, Campbell, Espin, and McMaster (2013) compared EL student writing performance on CBMs to their writing performance on several standardized measure. The standardized measures included the "Test of Written Language-III, the writing subtest of the Test of Emerging Academic English, and the Minnesota state writing test" (Campbell et al., 2013, p. 431). For high school students, Campbell and colleagues concluded that such CBMs are appropriate for use with EL students.

Because CBMs can be altered to ensure that the student understands the tasks presented, they have also been used to determine skill deficits among students from culturally- and linguistically-diverse backgrounds. For example, Baker and Good (1995) conducted a study of the technical adequacy of R-CBMs with Spanish EL and bilingual students. Over a 13-week period, 76 native English speakers, bilingual Hispanic students, and Spanish ELs were given an ORF and a Maze R-CBM in order to compare each measure's predictive ability on statewide assessments for both groups.

Through the use of standard R-CBM procedures provided by Shinn (1989), Baker and Good (1995) found a correlation between the state-wide tests and both ORF and Maze performance for all students. They also found that such fluency measures were equally sensitive to reading growth for bilingual and EL students as they were for native English speakers. By looking at the alternative-form reliability coefficients that fell above .80, the concurrent construct validity coefficient between .51 and .82, and the discriminatory construct validity of the CBMs with English proficiency measures, Baker and Good (1995) concluded that ORF R-CBMs are equally reliable and valid for bilingual, EL, and native English speakers.

Similarly, Shapiro et al. (2006) studied the relationships between R-CBMs, Math CBMs, and statewide assessments. More specifically, they looked at how R-CBMs, math computation

and concept application skill probes, the Pennsylvania System of School Assessment (2019), and several other standardized assessments correlated with one another. Overall, Shapiro and colleagues found "moderate to strong correlations" between the CBM probes and the standardized tests. While the study did not focus primarily on ELs, Shapiro et al. (2006) noted that one of the two districts that participated had a "relatively high" percentage of EL students in their sample (p. 32).

As noted previously, R-CBMs can also be used as pre- and post-test measures to assess instructional practices. Wayman et al. (2010) studied the ability of CBMs to measure the effectiveness of peer-mediated instruction with EL students. More specifically, Wayman and colleagues wanted to assess whether CBMs were an effective pre- and post-test measure of EL student academic growth in determining the effectiveness of a peer-mediated program known as Peer-to-Peer Adapted Layered Streaming (Rejaie & Ortega, 2003). Overall, the study concluded that CBMs could accurately measure the academic impact of instructional practices used with EL students, allowing teachers to better fit the needs of their students.

CBMs are useful in helping EL students improve their early literacy acquisition. Ergul (2007) studied the effectiveness of teachers making decisions based on CBMs. In the study, CBMs supplemented regular instruction by incorporating weekly assessments, instructional modifications based on data, and evaluations of the outcomes. The newly modified CBM was called Curriculum-Based Decision Making (CBDM) and was intended to enhance teachers' instructional planning and the students' learning. Ergul (2007) divided 291 EL preschoolers into three groups: a CBM-only group, a CBDM group, and a control group that did not incorporate any CBMs into their instruction. All students were then further divided into high- and low-risk groups. Overall, the preschoolers placed within the CBDM group showed significantly greater

improvements on their early literacy posttest; these effects were observed in both the high- and low-risk EL students. Ergul (2007) concluded that using a structured evaluation of CBM data to make informed decisions about ELs in the classroom could significantly effect their early literacy skills acquisition.

A study conducted by Graves et al. (2005) looked at first-grade students' reading progress by assessing ORF and Nonsense Word Fluency (NWF) gains for both EL students and native English-speakers. NWF tasks require students to read as many words from a list of fake words as quickly and as accurately as possible. Performance was measured according to the total of correct nonsense words read. Similarly to Baker and Good (1995), Graves et al. (2005) found that EL and native-English speakers made similar gains in word accuracy rates across the six-week assessment period. In addition, Graves et al. (2005) found positive but weak correlations between both R-CBM measures and language proficiency ratings for EL students. These findings suggest that, while CBMs could be used reliably for both EL and native English speakers, language skills prior to entering first-grade "were not a good predictor of how well they would read at the end of first grade" (Graves et al., 2005, p. 223).

Espin et al. (2012) elaborated on these findings through a district-wide report about normative assessment of EL students. In a case sample, Espin et al. (2012) reported data collection involving EL students' performance on math (i.e., math fact probe), written expression (i.e., story starters), and reading (i.e., ORF and Maze probes) CBMs across twenty different elementary schools in the Saint Paul public school district. Overall, EL student performance in the fall, winter, and spring CBM administrations showed strong consistencies when compared to native English speaker performances**.** The EL students' growth scores were also similar to the

national sample, supporting the reliability and validity of such measures for EL students (Espin et al., 2012).

When thinking about EL students and their English language development, consideration of students' proficiency levels in their native language is essential. Laija-Rodríguez et al. (2006) conducted a study that investigated the relationship between English and Spanish CALP levels. Spanish-speaking EL students in the second and third grade participated in the study. CALP levels were measured using the Woodcock-Muñoz Language Survey (WMLS; Woodcock, 2005) and reading growth was measured by ORF R-CBMs. Overall, Laija-Rodríguez and colleagues found a weak but significant relationship between Spanish and English CALP levels. A significant, but weak relationship was also found between reading growth in both languages. These findings suggest that language proficiency in EL students' native language may impact their language attainment in English (Laija-Rodríguez et al., 2006; Ordóñez et al., 2002).

Keller-Margulis and Mercer (2014) also conducted a study that assessed how English R-CBM performance related to Spanish R-CBM performance. More specifically, they looked at both initial benchmark scores and growth in both languages to compare students' reading skills. Overall, Keller-Margulis and Mercer (2014) found that performance and growth were strongly correlated across both languages for grades one through five. These findings emphasize the importance of incorporating the student's native language into instruction to help further facilitate English language proficiency. In addition, these findings support the use of CBMs as a simple and brief method to dual-language assessment.

Predictive Ability of Curriculum-Based Measures and EL Students

CBMs have been shown to accurately predict academic outcomes for EL students (Codding et al., 2015; Kim et al., 2016; Muyskens, Betts, Lau, & Marston, 2009; Ramírez et al.,

2007; Sandberg & Reschly, 2011; Shapiro et al., 2006). While looking at the potential predictive abilities of CBMs, Codding et al. (2015) conducted a study that looked at the relationship between reading, mathematics, and writing CBMs with high-stakes assessment at the secondary level. More specifically, Codding et al. (2015) focused on the correlation between students' AIMSweb ORF, reading comprehension (i.e., MAZE), written expression, and math computation CBM performance to students' performance on the Massachusetts Comprehensive Assessment System (2018) for English Language Arts (MCAS-ELA) and Mathematics (MCAS-M). After assessing seventh-grade student performance on all measures, Codding et al. (2015) found that CBMs could reliably measure yearly growth and predict student performance on the MCAS-ELA and MCAS-M for EL students in middle school. Moreover, Codding et al. (2015) note that "reading was the strongest predictor" (p. 437) of student performance on the MCAS-ELA and showed moderate prediction for the MCAS-M. These findings support the importance of reading instruction and language proficiency, showing how reading skills affect many academic areas (Codding et al., 2015).

Similarly, Kim et al. (2016) studied the predictive validity of R-CBMs for EL students. In their study, Kim and colleagues looked at the performance of Spanish-speaking EL students in the third grade to assess the relation between both DIBELS Oral Reading Fluency (DORF) and Daze probes and the California Standards Tests–English Language Arts (CST–ELA), a statewide reading assessment. The DORF task is the same as ORF probes, measuring a student's oral reading fluency and accuracy, and Daze probes are the same as Maze probes, measuring a student's reading comprehension skills (Smith & Wallin, 2011). Overall, Kim et al. (2016) found a moderate to strong relationships between the R-CBM probes and the CST-ELA. Results also indicated that while both CBM probes significantly predicted CST-ELA performance, the DORF

probe had larger effect sizes, suggesting that the DORF task "was a stronger predictor of reading outcomes" (Kim et al., 2016, p. 1).

Muyskens et al. (2009) also examined the concurrent and predictive validity of R-CBMs with EL students. In their study, Muyskens and colleagues used ORF probes as a predictor of reading achievement in fifth grade EL students. Standardized reading achievement was measured by a statewide reading assessment and compared to previous performance on the ORF R-CBM. In accordance with the results presented by both Codding et al. (2015) and Kim et al., (2016), Muyskens et al. (2009) concluded that ORF R-CBMs are a significant predictor of later reading achievement in EL students. Stokes (2010) broadened these findings by adding in English proficiency levels and R-CBM growth rates to the analysis. More specifically, Stokes (2010) studied the predictive validity of ORF R-CBMs with sixth graders by comparing their performance to their reading achievement on a high-stakes assessment. The average rate of growth on the R-CBM probes was also assessed and compared to the students' level of English proficiency. Of the 350 students in the sample, 90 were characterized as EL students with varying levels of proficiency. Stokes (2010) found that, while EL students significantly performed lower than their native English peers initially, their growth trajectories did not vary. Additionally, Stokes (2010) concluded that initial R-CBM performance, R-CBM growth rates between the fall and spring, EL status, and English language proficiency were all significant predictors for future performance on standardized achievement measures.

Similar to the study conducted by Stokes (2010), Farmer (2013) expanded on the previous studies by looking at whether ORF R-CBM probes were effective tools for assessing EL student growth and for identifying EL students who were academically at risk of failing the standardized state exams. Overall, the study concluded that R-CBMs were effective in measuring

both academic growth and in identifying EL students who were at risk. Jimerson, Hong, Stage, and Gerber (2013) also examined ORF growth in EL students. In their study, ORF growth and socioeconomic status were used to predict EL student performance on the Stanford Achievement Test – Ninth Edition (SAT-9). The SAT-9 is a measure of overall achievement, focusing on reading and mathematics performance (Jimerson et al., 2013). Results indicated that both variables reliably predicted SAT-9 performance for students between the first and fourth grades. Similar to the findings presented by Stokes (2010), Jimerson et al. (2013) found that reading fluency trajectories for EL students were not significantly different from the trajectories of native English speakers.

To further assess language development for EL students, Ramírez et al. (2007) looked at the relation between a child's ORF in English and his/her ORF in Spanish. In this study, 68 Spanish-speaking EL children between first and fifth grade were assessed in the fall, winter, and spring using the ORF R-CBMs in both languages. The students' performance on the CBM tasks were also compared to their Texas Assessment Academic Skills reading performance and the Developmental Reading Assessment performance to assess the predictive ability of both ORF R-CBMs (Ramírez et al., 2007). With the exception of grade four, Ramírez et al. (2007) found that both English and Spanish ORF R-CBMs were sensitive to growth across the other grade levels and were significantly correlated to one another with a moderate correlation (i.e., .71-.79). Ramírez et al. (2007) also found that scores on the Spanish ORF R-CBM "significantly predicted English reading outcomes" (p. 795). One issue noted by Ramírez et al. (2007) was that student performances on the English ORF probes were consistently less accurate for all grade levels when compared to their performance on the Spanish ORF probes. This lower accuracy for the English ORF R-CBMs may indicate that reading "accuracy and speed did not emerge

concurrently in the two languages" (Ramírez et al., 2007, p. 801). From the conclusion of various studies, it is evident that CBMs provide a valid and reliable method of assessing EL student skills and monitoring their academic progress (Codding et al., 2015; Farmer, 2013; Jimerson et al., 2013; Kim et al., 2016; Ramírez et al., 2007; Stokes, 2010).

Disadvantages of Curriculum-Based Measures

While a substantial amount of data supports the use and utility of CBMs in an educational setting, educators should be aware of certain issues when using such measures. Rhodes et al. (2005) note some disadvantages of using CBM probes. The first is the variability of CBM materials across teachers. Such variability can restrict comparisons across classrooms by using measures that assess different skills. Another issue noted by Rhodes et al. (2005) is the heavy reliance on visual data to make educational decisions. By depending on visual data, source error may increase, leading to inaccurate decisions being made about the child's academic needs. Furthermore, Rhodes et al. (2005) warns that teachers need to be careful not to teach materials that specifically target the CBM probes. In other words, teachers must be mindful as to whether or not they may be teaching to the test or the results may provide an inaccurate reading of a child's actual skills.

Within the last two decades, many different materials have been published for R-CBMs. Given this expansion, Shapiro (2011) argues that the types of materials being used have shifted and may not reflect current grade-based reading materials. In order to ensure that the current CBMs are valid, grade-level readability of each selected passage chosen should be determined (Shapiro, 2011). Another issue presented by R-CBMs is the methods used by different studies to support their use. Performance and growth on R-CBM probes have heavily depended on timed samples of ORF (Deno, 2003; Moore, 1997). More specifically, R-CBM probes have focused on

ORF performance to make educational decisions on a student's current reading skills (Deno, 2003). Because of this limited scope, future research should broaden the types of R-CBMs studied and reassess the reliability of CBMs as a predictive measure.

While some studies have found strong support for the use of R-CBMs to measure EL student growth and predict future performance on standardized assessments, several others have found contradictory results (Baker et al., 2015; Farmer, 2013). For example, Keller-Margulis, Clemens, Im, Kwok, and Booth (2012) conducted a study on third, fourth, and fifth graders to test whether R-CBMs can be used as an accurate measure of academic reading growth. The study consisted of both native English speakers and EL students. Overall, the study concluded that while reading growth was accurately measured for native English speakers in the third and fourth grade, R-CBMs were not able to consistently assess reading growth for EL students nor students in the fifth grade.

Likewise, Baker et al. (2015) studied the criterion validity of R-CBMs and their ability to predict seventh- and eighth-grade performance on a statewide assessment. More specifically, they looked at the individual predictive abilities of ORF, reading comprehension, and word reading accuracy tasks and how well they can predict student performance on the standardized reading exam. Overall, Baker and colleagues found that the results were inconsistent for EL students. For native English speakers, 55-58% of the state-wide assessment could be explained by the ORF and reading comprehension CBMs. Conversely, the specificity rate for EL students was low and resulted in an inaccurate prediction of how the student would perform on the standardized measure. Overall, Baker et al. (2015) concluded that educators should use ORF probes with caution when assessing EL students in the seventh and eighth grade. This conclusion

contradicts previous research stating that R-CBMs are "equally effective predictors for ELs and non-ELs" (Baker et al., 2015, p. 98).

One issue that may impact the validity of using CBMs with EL students is the equivalence of scores (Vanderwood et al., 2014). Equivalence refers to the measure's ability to accurately assess individuals of different groups, such as native English speakers and EL students. Because CBMs were normed primarily on a sample of native English speakers, Vanderwood et al. (2014) argue that CBMs may not have accurate scores for EL students. Keller-Margulis and Mercer (2014) add, "it is becoming increasingly apparent that national norms are not always appropriate for diverse groups of students" (p. 689). Language differences between the student's native language and English may strongly impact their performance on these measures (Spinelli, 2008; Vanderwood et al., 2014). As such, CBMs may be biased to students of culturally or linguistically diverse backgrounds.

Richardson, Hawken, and Kircher (2012) assessed whether R-CBMs were equally predictive for EL students who primarily spoke Spanish at home as they were for native English speakers and students of diverse backgrounds. Students in the third, fourth and fifth grades participated in the study and were compared to their non-EL peers. Using hierarchical linear modeling on Maze R-CBM scores, Richardson et al. (2012) found that an intercept bias existed for both EL students and culturally diverse backgrounds. These findings suggest that while R-CBMs are strong predictors for future academic performance, their predictive abilities may be limited for those of culturally- or linguistically-diverse backgrounds.

Another issue that may differentially affect EL students when using CBM probes is the variability of scores between raters. While all CBM suppliers provide instructions on how to score R-CBM passages, providers rarely delineate clear instructions on how to score language

differences presented by students with limited English proficiency. For example, in the ORF scoring instructions from Intervention Central, Wright (2013) states that any "examples of dialectical speech are counted as correct" (p. 3). While these instructions seem clear, Wright (2013) goes on to state that any "mispronunciations are counted as errors" (p. 3). Similarly for AIMSweb, Shinn and Shinn (2002) note that dialectical speech does not count as an error but that mispronunciations do count as errors when scoring CBMs. Differences in dialect and associated mispronunciations of words can be extremely difficult to distinguish when assessing EL students with unique accents. According to Derwing and Munro (2005), this difficulty can lead teachers and other educators "to rely on own intuitions with little direction" (p. 379). While a student's accent can lead one rater to consider the pronunciation a dialectical norm, another rater may consider the difference a mispronunciation, leading to vastly different ratings of the same student on the same R-CBM passage.

Interrater discrepancies can affect educational decisions made for the student, and such discrepancies may even affect the student's eligibility for special education. While a great deal of research has looked at the predictive ability and utility of R-CBMs with EL students, little to no research has focused on the IRR of ORF R-CBM probes with EL students. Particularly given that EL students often speak with unique accents, parsing out errors due to mispronunciations of words and differences in dialect is challenging for raters. If scored consistently, ORF R-CBMs can be a useful tool when assessing EL students.

Interrater Reliability

Interrater reliability (IRR) or interobserver agreement has been defined as the consistency of observations between raters or observers (Bryington, Palmer, & Watkins, 2002; Suen, 1988). IRR aims to assess the degree to which two or more observers agree or disagree on the

occurrence or nonoccurrence of a behavior (Bryington et al., 1988; Hintze, 2005; Suen, 1988). If the IRR for a measure is high, researchers can infer that the ratings are more consistent between observers. As such, it is assumed that "data from any one of the observers is relatively free from measurement error" (Suen, 1988, p. 349).

Various methods can be applied to assess IRR (Bryington et al., 2002). While all of the statistical methods assess observer rating similarities and produce reliability coefficients, each method differs in its treatment of chance agreement and how many observers can be included (Suen, 1988). The simplest and most common method is known as percent agreement (Bryington et al., 2002; McDermott, 1988). For this method, ordinal data between two raters are compared to produce a percent agreement (Hintze, 2005; McDermott, 1988). Percent agreement evaluates IRR without correcting for the possibility that the results may have occurred due to chance (Suen, 1988). Because IRR is calculated without any corrections, it can be difficult to make interpretations and to ensure that the results were not due to chance (Bryington et al., 2002). Unfortunately, percent agreement is one of the most common methods for evaluating IRR for CBMs (Ardoin, Roof, Klubnick, & Carfolite, 2008; Gansle, VanDerHeyden, Noell, Resetar, & Williams, 2006; Lewandowski, Codding, Kleinmann, & Tucker, 2003; Yeo, 2011).

Pearson's *r* is also a very common method used to assess IRR between two observations that are interval or ratio in nature (McDermott, 1988). The *phi* coefficient is similar to Pearson's *r* but is used when data is dichotomous (Hintze, 2005; McDermott, 1988; Suen, 1988). Both methods provide the degree of variance that is due to random observer error and work best when assessing the degree of agreement among standardized scores. According to Suen (1988), however, Pearson's *r* and the *phi* coefficient "do not take into account systematic differences between the two observers and random observer errors are within-observer in nature" (p. 352).

Two additional common statistical methods used to assess IRR are chi-squared and occurrence/nonoccurrence agreement (McDermott, 1988). While other methods look at agreement between two raters, the chi-square method focuses on association between scores. As such, it is often only used to test the relationship between data types and categories and not agreement of outcome data (McDermott, 1988). Occurrence/nonoccurrence agreement partially corrects for chance agreement (Hintze, 2005; Suen, 1988). While this method looks at the percentage of time that observers agreed, it still does not fully control for chance and can only be calculated with dichotomous data from two observers (Bryington et al., 2002).

Cohen's Kappa (*K*) is a more comprehensive method for assessing IRR by statistically removing chance agreement from the estimation (Bryington et al., 2002; Hallgren, 2012; Hintze, 2005; Suen, 1988). *K* produces proportional data (i.e., scores between -1 and 1) instead of percentages to evaluate agreement through the sums of squares produced (McDermott, 1988). In addition, the standard error is calculated to determine result significance. While *K* does analyze proportion of agreement beyond chance, *K* can only be used with two observers (Hallgren, 2012; McDermott, 1988). Additionally, *K* is also limited to data that is dichotomous or nominal (Bryington et al., 2002). To correct some of these issues, variations of *K* have been produced. Kappa *M*, Light's *G*, and Fleiss's *K* are able to analyze data from multiple observers (Bryington et al., 2002; McDermott, 1988). However, it should be noted that all the *K* variations are still limited to nominal data (Hallgren, 2012; McDermott, 1988).

While the above methods may be useful in understanding overall reliability of data collected, most ignore the possible contributions made by different sources of error, are limited to two observations, or are only used with nominal or categorical data (Hallgren, 2012; Suen, 1988). Alternatively, the intraclass correlation (ICC) method is able to assess both overall

reliability and additional error present in the data (Bujang & Baharum, 2017; Suen, 1988). In ICC, analysis of variance (ANOVA) is used in the error analysis to determine the magnitude of error present from different sources (Hintze, 2005). According to Suen (1988), results obtained from ICC can help "guide the improvement of the observation situation" (p. 353). As such, Suen (1988) states that ICC can be used to determine the reliability of the scores for both normative and criterion-referenced interpretation, as well as to assess the average scores between multiple subjects and between multiple groups of raters.

ICC can be used with ordinal, interval, and ratio data and produces mean squares across both subjects and raters (Hallgren, 2012; Suen, 1988). Additionally, ICC produces residual mean squares. Through further analysis, these statistical outputs can be used to assess the random error variance, variance due to raters, and variance due to subjects (Hintze, 2005; Suen, 1988). Overall, the ICC method is the most appropriate for evaluating IRR of ORF R-CBMs that include both multiple raters and multiple subjects (Bujang & Baharum, 2017; Hallgren, 2012).

CHAPTER 3

Method

Participants

Spanish is the most predominant language spoken by EL students in the U.S. (McFarland et al., 2017; Sandberg & Reschly, 2011). The largest number of EL students in the U.S. is currently found between kindergarten and fourth grade (Keller-Margulis & Mercer, 2014; McFarland et al., 2017). Because kindergarten R-CBM probes do not include ORF passages, kindergarteners will not be included in the study. Over 20 Spanish-speaking countries exist around the world. In order to incorporate multiple Spanish dialects into the analysis, speakers from different Spanish-speaking countries were preferred within each grade level. A total of eight Spanish-speaking EL student speakers between the first and fourth grade were recruited. More specifically, one female and one male student participant were selected for each grade level. Speakers were recruited through a local suburban public school's EL program in Pennsylvania. To increase student speaker participation, all students' parents opted to keep personal information anonymous for this study.

According to the U.S. National Center of Educational Statistics (NCES, 2012) School and Staffing Survey (SASS) data from the 2011-2012 school year, approximately 41,880 school psychologists were working in the United States. Of those employed, only 14% were fluent in another language and less than 8% of bilingual school psychologists serviced students in a language other than English (National Association of School Psychologists [NASP], 2017). A total of 198 NASP-approved school psychology programs across the U.S. were contacted directly and asked to participate. Study information was also posted on several discussion boards

to increase participation, including the NASP, New York Association of School Psychologists (NYASP), and Association of School Psychologist of Pennsylvania (ASPP) discussion boards.

According to Bujang and Baharum (2017), ICC calculations do not usually require a large sample size to determine level of agreement, especially when the total number of observations made by each rater increases. Additionally, Koo and Li (2016) suggest a minimum of 30 samples for a minimum of 3 raters, resulting in 90 ratings. Overall, a total of 65 practicing school psychologists and/ or current school psychology students contacted the researcher to participate. Of these, 35 total participants completed both Time 1 (T1) and Time 2 (T2) of the study. Each rater scored a total of eight EL passages, four during T1 and four during T2, resulting in a total of 35 ratings for each administration.

Of the 35 raters, 32 (91.43%) were female. The majority of raters held a bachelor's degree (30.56%) with lower proportions for raters with master's degrees (25%), doctorate degrees (22.22%), educational specialists (11.11%), and other qualifications (11.11%). Most raters were currently students in a doctoral-level school psychology program (47.22%). The remainder of raters were currently working as school psychologists in the field (30.56%), faculty members at a school psychology training program (11.11%), students in a master's school psychology program (2.78%), or an individual completing postdoctoral research (2.78%). Approximately 5.56% raters marked Other when asked about their current standing in the field of school psychology. The majority of faculty members, school psychologists working in the field, and Postdoctoral student raters indicated that they have been working in the school psychology field for less than three years (50%). About 27.78% have been working in the field for 4 to 10 years, 11.11% between 11 and 19 years, and 11.11% for more than 20 years.

Overall, the majority of raters (44.44%) indicated that they had novice Spanish proficiency (i.e., able to read, write, and/or speak Spanish with extensive assistance). A total of 16.67% had intermediate Spanish proficiency (i.e., able to read, write, and speak Spanish with some assistance), 27.78% had low Spanish proficiency (i.e., unable to read, write, or speak Spanish), and 11.11% had advanced proficiency (i.e., easily able to read, write, and speak Spanish without assistance). Approximately 66.67% of raters had received formal Spanish instruction. Of these raters, 47.83% received one to two years of college-level Spanish courses, 43.48% only received high school Spanish instruction, and 8.7% majored in Spanish as an undergraduate. Additionally, 25% of all raters indicated that they spoke another language, including Japanese, French, German, Italian, and American Sign Language.

Measures

The ORF R-CBM probes were obtained from AIMSweb, a commonly used and reliable online source for CBMs. The ORF R-CBM probes require the student to read a passage aloud for 1 minute while the rater marks any errors made by the student according to the guidelines provided by AIMSweb (Shinn & Shinn, 2002). Errors include omitted words, substituted words, words pronounced incorrectly, and any words supplied to the student (Shinn & Shinn, 2002). The ORF probes provide both reading accuracy and reading fluency scores. Reading accuracy is presented as a percentage calculated by dividing the total number of words read correctly by the total number of words attempted. Reading fluency refers to the total number of words read correctly in one minute. Previous research on AIMSweb ORF R-CBM scores have established adequate reliability and validity coefficients. Validity coefficients were found to typically be in the .60 - .80 range when compared to other established criterion measures. Test-retest reliability range estimates were found between .89 and .94, and IRR fell around .99 (Shinn & Shinn, 2002,

p. 35). Grade-level readability of each R-CBM probe were evaluated using the Microsoft Word (Version 16.31) readability index (i.e., the Flesch-Kincaid Grade Level test) to verify that each passage was appropriate for the reader. The Flesch-Kincaid Grade Level test looks at both ASL (i.e., number of words divided by the number of sentences) and ASW (i.e., number of syllables divided by the number of words) to calculate the passages' grade level readability statistic (Microsoft, 2018). This ensured that the selected passages read were appropriate for the elementary school-aged readers.

Procedure

Parents of student readers were asked to complete a short survey before participating in the study. The survey included information on their native country, the number of years in the U.S., whether English is spoken in the home, and what grade they are currently in. Surveys were optional to increase student speaker participation. All parents opted to not complete the survey in order to remain anonymous. As such, only the student's current grade level and gender were available. In addition, parents were verbally asked about any disabilities that may impact their students' reading and speaking abilities (i.e., hearing, voice impairment, vision, etc.). Speakers with additional disabilities were excluded from the study to avoid conflicting variables and factors. Audio recordings were done using a handheld audio recorder as well as the researcher's personal laptop recorder to ensure the clearest recording was used. Each ORF passage was then coded according to the student's grade and gender to ensure anonymity of the speakers.

All 35 raters were randomly assigned to Group A or Group B, resulting in 17-18 raters per group. Raters within the same group received the same eight passages to score, four passages during T1 and four passages during T2. Given that eight EL passages were available, each

passage received 35 ratings (i.e., 17 to 18 ratings for T1 and 17 to 18 ratings for T2). This resulted in a combined total of 70 ratings (i.e., 35 ratings for T1 and 35 for T2).

All raters received the audio recording of the four Spanish-speaking EL student speakers reading ORF passage they were assigned through a Qualtrics (Version Aug. 2019) form along with a short survey asking for their experience levels in the school psychology field. Such survey questions included the years spent as a school psychologist or years of training within their program, the extent of experience working with EL students, and whether they speak Spanish or have taken any courses to learn Spanish. Raters heard the recordings on a personal device that has audio capabilities (i.e., smart phones, laptops, etc.). The T1 instructions, directions, and scoring procedures created by AIMSweb were supplied to each rater without any further instruction on how to score words that may have been pronounced differently due to speakers' accents (Pearson, 2012). Raters were instructed to follow the directions and score each ORF R-CBM according to the AIMSweb instructions (see Appendix A). Upon completion, raters then returned their scored R-CBM passages and notations back to the researcher via Qualtrics. The fluency and accuracy scores generated for the probes were then used to compare ratings of the speakers' ORF skills. In addition, each word within the passage was coded dichotomously according to the rater's notations (i.e., *correct* = 1; *incorrect* = 0) and compared to other ratings of the same passage. The errors that were consistently identified by a majority of raters among all eight passages for T1 were then used to help create the T2 administration guidelines (see Appendix B).

As per the new instructions, words read with rolled "r" sounds, words read without ending sounds (e.g., -s, -ed, etc.) that did not change the word's meaning (e.g., "sleeping" said as "sleepin "), and words with "th" sound replaced with "d" sound were to be marked as correct.

Additionally, words read with inserted sounds that did change the word's meaning (e.g., "trees" instead of "tree") were to be marked as incorrect. Approximately two months later (T2), the raters received the four remaining audio recordings and passages with the more detailed set of instructions on how to handle words said incorrectly due to the speakers' accents. Once the R-CBM probes were scored, raters returned the completed R-CBM probes to the researcher for analysis.

Design and Analyses

Descriptive statistics were used to describe the sample demographics and the distribution of accuracy ratings for each ORF passage for both administrations (e.g., *M*, *SD*, range). Qualitative data also assisted in determining where raters agreed the most and where their responses differed. Error analyses from T1 discrepancies were used to create enhanced ORF R-CBM guidelines for scoring passages read by EL students. The enhanced ORF R-CBM instructions provided to raters for the T2 administration of the passages were created using the anecdotal and observational data collected from T1.

When selecting ICC as the statistical method of a study, researchers must identify the model (i.e., one-way or two-way), the type (i.e., mean of ratings or single rating comparison), and the definition (i.e., absolute agreement or consistency) of the planned analysis (Kim, 2013; Koo & Li, 2016; McGraw & Wong, 1996; Shrout & Fleiss, 1979). Given that the current study aimed to compare ratings from multiple raters of the same sample drawn from a larger population, a two-way random-effects ICC statistical model was the most appropriate method to use for T1 and T2 (Hallgren, 2012; Koo & Li, 2016; McGraw & Wong, 1996; Shrout & Fleiss, 1979). More specifically, the current study anticipated that individual differences would impact ratings for each passage, thus comparisons were made according to the mean ratings for each

group in T1 and in T2 (Koo & Li, 2016; Shrout & Fleiss, 1979). Further, given the focus of the current study on how reliable raters were with each other rather than how consistent their scores were between subjects or over time, the absolute agreement definition of ICC was utilized to analyze ratings for each group (Kim, 2013; Koo & Li, 2016). Lastly, to compare the ratings for passages in T1 to the ratings for the corresponding passages in T2, each passage's comparison was tested for significance using the two-way mixed-effects model (Koo & Li, 2016). Absolute agreement procedures were used to assess the differences in ratings between administrations (Koo & Li, 2016). Overall, all analyses were calculated using SPSS software (Version 26.0).

CHAPTER 4

Results

Descriptive Statistics

Means (*M*), standard deviations (*SD*), and ranges for the accuracy ratings for each passage for both T1 and T2 are reported in Table 1. As previously noted, accuracy rates are presented as a percentage calculated by dividing the total number of words read correctly by the total number of words attempted. Overall, accuracy rates of each individual passage remained within 6 percentage points between T1 and T2.

Table 1

Means, Standard Deviations, and Ranges of Each Passage for Time 1 and Time 2

Passage	Time 1 (T1)			Time 2 (T2)		
	M	*SD*	Range	*M*	*SD*	Range
Passage 1	95.71%	1.55	92.73% - 99.10%	96.82%	1.37	93.64% - 98.18%
Passage 2	64.70%	6.10	52.63% - 73.68%	65.33%	4.94	52.63% - 73.68%
Passage 3	42.81%	6.32	29.41% - 52.94%	48.04%	6.77	35.29% - 52.94%
Passage 4	95.33%	2.31	91.18% - 99.02%	95.04%	1.71	90.20% - 96.08%
Passage 5	90.30%	3.43	85.00% - 95.00%	91.11%	2.06	86.67% - 95.00%
Passage 6	61.15%	5.80	50.00% - 72.92%	59.80%	3.95	52.08% - 68.75%
Passage 7	57.09%	6.81	48.15% - 70.37%	59.74%	6.06	46.30% - 68.52%
Passage 8	58.61%	8.42	42.59% - 74.07%	55.01%	5.96	50.00% - 68.52%

Note. Accuracy rates are presented as a percentage calculated by dividing the total number of words read correctly by the total number of words attempted.

Comparison within Time 1 and Time 2

A two-way random-effects ICC statistical model was used to compare ratings from multiple raters of the same sample pulled from a larger population. The ICCs and their 95% confidence intervals for T1 and T2 by group are reported in Table 2. Overall, all ICC values

were statistically significant at $p = .05$ level. According to Koo and Li (2016), absolute ICC "values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability" (p. 158). Both groups indicated excellent reliability for both T1 and T2.

Table 2

Intraclass Correlations Coefficients and 95% Confidence Intervals for Time 1 and Time 2 by Group

Group	Time 1 (T1)		Time 2 (T2)	
	ICC	CI	ICC	CI
Group A	.998*	[.993 - 1.00]	.996*	[.988 - .1.00]
Group B	.993*	[.974 - .999]	.998*	[.993 - 1.00]

Note. ICC = Intraclass Correlation Coefficient; CI = 95% Confidence Interval.

*$p < .05$.

Comparisons between Time 1 and Time 2

A two-way mixed-effects model was used to compare the accuracy ratings for passages in T1 to the accuracy ratings for the corresponding passages in T2. The ICCs and 95% confidence intervals between T1 and T2 are reported in Table 3. According to Koo and Li (2016) ICC guidelines, all passages in T1 and T2 indicated excellent reliability. Additionally, a two-way mixed-effects model was used to assess the consistency of *M* ratings for each passage between T1 and T2 (see Table 3). Overall, ratings between T1 and T2 remained consistently reliable over time, regardless of the instructions provided. All ICCs produced were statistically significant at the $p < .05$ level.

Table 3

Intraclass Correlations Coefficients and 95% Confidence Intervals Between Time 1 and Time 2

Time	ICC	CI
Time 1 (All Passages)	.997*	[.99 - 1.00]
Time 2 (All Passages)	.994*	[.99 - 1.00]
Consistency Between Time 1 and Time 2		
Time 1 v. Time 2	.996*	[.978 - 1.00]

Note. ICC = Intraclass Correlation Coefficient; CI = 95% Confidence Interval.

*$p < .05$.

CHAPTER 5

Discussion

The purpose of the current study was to examine the IRR of ORF R-CBMs when used with Spanish-speaking EL students. The primary research questions were to assess the degree to which raters differed from one another when scoring the ORF R-CBM passages and to assess the change in IRR when raters were given a more detailed set of scoring guidelines. Previous research has consistently supported the use of ORF R-CBM passages with EL students (Baker & Good; 1995; Domínguez de Ramírez & Shapiro, 2006; Fuchs, 2016; Gersten et al., 2007; Wiley & Deno, 2005). Similarly, the current results suggest that, regardless of which scoring guidelines were provided, all EL passages consistently had excellent IRR.

While these results do not support the initial hypothesis that raters would differ initially when scoring the passages, alignments between raters for all EL student passages is a positive outcome overall, especially given the impact that CBMs can have on EL student educational experiences and long-term outcomes (Codding et al., 2015; Ergul, 2007; Kim et al., 2016; Ramírez et al., 2007). These results may be due to the established strong psychometric properties of R-CBMs from reputable and widely-used providers, such AIMSweb (Shinn & Shinn, 2002). Another explanation for the current findings is that the current study's limitations may have impacted the results.

Limitations

A number of limitations presented in the current study. The current sample of EL student speakers was small and consisted of students from only one ESL program from a suburban public school in Pennsylvania. Due to the limited number of EL passages, the variability of Spanish dialects between EL students may have resulted in skewed results. In addition, a small

sample size of raters was recruited from school psychologists that were either directly connected to a university or were part of a professional organization of school psychologists (e.g., ASPP or NYASP). Both the university and the professional organizations typically emphasize the importance of best practices, thus may result in higher rates of IRR than those typically seen in the overall population of school psychologists. As such, results may have been impacted by both the small sample of EL student speakers and the small sample of raters. Koo and Li (2016) note that having a small number of subjects or a small number of raters may result in a lack of variability among both subjects and raters, producing unreliable IRR results. It is recommended that a replication be conducted with a larger sample size of EL students from a more diverse Spanish-speaking population. In addition, replication studies should include a broader sample of raters from a larger population of school psychologists to help prevent skewed data and reduce the margin of error.

Current results suggest that, regardless of the instructions provided, raters demonstrated excellent reliability. One explanation for these results is that raters may not have followed the T2 set of directions as anticipated. Raters may have habitually followed the standard scoring procedures, making the IRR results from T2 less reliable. Without ensuring that raters used these instructions, it is difficult to make definitive statements about T2 results. Thus, future studies should require raters to comment on their views of the T2 instructions and further explain their ratings so as to help ensure the integrity of ratings and the use of T2 instructions for T2 passages.

Future Research

While much research has been done on the utility of R-CBMs, more research needs to focus on the reliability and validity of R-CBM ratings to predict overall EL student outcomes. Considering the current study results, a larger sample of Spanish-speaking EL students may

produce more variability within the sample of passages assessed. Including a more diverse population of Spanish-speakers from different Spanish-speaking countries with a variety of dialects may also be useful. Because similarities between the English language and other languages can vary, assessment of EL students from other countries and cultures would provide additional perspectives and understanding of R-CBM reliability for students of other backgrounds.

EL students' acquisition of English (i.e., BICS and CALP) might also influence performance on R-CBM measures. An EL student with high acquisition is likely to have less errors due to an accent than a student with weaker English acquisition, possibly resulting in higher IRR for the high-acquisition student. Having parents complete a language survey or participate in an interview might help determine a student's BICS and CALP levels. As such, controlling for an EL student's BICS and CALP levels may help further assess IRR of R-CBMs.

Overall, requiring the parents of the participating EL students to complete a demographic survey or interview may provide a better understanding of what factors might impact a student's R-CBM scores and general outcomes. The EL student survey or interview should include questions regarding the student's gender, current grade level, preferred language, country of origin, number of years in the U.S., and English instruction received. Additionally, existing cultural difference may impact a student's performance. The use of both parent input surveys, parent interviews, and acculturation measures are considered to be best practices when working with EL students to ensure a comprehensive assessment of student strengths and difficulties (Collier, 1998). Future studies should assess the impact of acculturation on R-CBM performance for EL students by providing acculturation forms to students as part of the study. The inclusion of the demographic survey or interview and the acculturation forms are likely to require that

parents disclose personal information to researchers, so particular caution should be taken if assessing the impacts of students' backgrounds on R-CBM performance.

Given the limited scope of the present study, it is also possible that data may have been skewed considering that all raters were recruited from NASP-approved school psychology training programs or from membership within two state-level school psychology associations (i.e., NYASP and ASPP). Involvement with these associations and direct involvement with the universities may have impacted the school psychologists' ratings due to a heavier emphasis on best practices. Recruiting a larger and more diverse sample of raters may help eliminate potential biases produced by participation in these organizations.

Overall, 72.22% of raters in the current study indicated that they had novice to advanced Spanish proficiency. Of these, approximately 66.67% had received some level of formal Spanish instruction. Rater exposure to Spanish language instruction prior to the study, as well as an increased level of familiarity with different accents and regional dialects of Spanish may have impacted the IRR of the R-CBM ratings. With more Spanish education and exposure, raters may have had a more thorough understanding of Spanish accents in everyday speech, resulting in higher IRR. As such, future studies should include raters that have not had any Spanish language instruction and reassess the reliability of R-CBM probes with Spanish EL students.

When R-CBMs are used in schools, they are commonly given by teachers and reading specialists. Given that the current study focused on IRR of school psychologists scoring R-CBMs, future studies should include a broader variety of raters to ensure that other practitioners in the education field are also scoring these measures reliably. In addition, the current study allowed raters to play the audio recording of each EL student passage as often as they liked to ensure that raters were being cautions with their scoring. Due to the brief nature of R-CBMs, it is

not common practice to record and review the student reading each passage. In schools, students are typically pulled from their instruction and given the brief R-CBM probe while teachers score their performance concurrently. As such, future research should assess whether allowing a rater to repeatedly listen to an R-CBM might impact the IRR of such passages.

As previously mentioned, there are many ways to interpret IRR. Various statistical analyses could have been used to interpret IRR results. Given the aim and purpose of the current study, the IRR interpretation of ICC as presented by Koo and Li (2016) was believed to be the best method to explain the IRR of the R-CBM passages read by the eight EL students. Future studies should aim to use other statistical means of interpreting the IRR of these passages to provide a variability in methods used to support or contradict the current findings.

Finally, several authors note that one issue R-CBMs pose for EL students is the inability to distinguish between EL students who are struggling due to their Limited English Proficiency (LEP) and EL students who are struggling due to underlying learning disabilities (Deno, 2003; Henderson et al., 1993; Klingner et al., 2008). Future research should assess the impact of LEP on the IRR of R-CBM ratings. Furthermore, future research should attempt to distinguish EL difficulties that are due to an underlying disability from those that are due to LEP.

Practical Implications

Performance on R-CBMs can significantly impact a student's educational outcomes. Concerns have been noted with EL students' underrepresentation in gifted programs, overrepresentation in special education for learning disabilities, higher rates of placement in least-restrictive environments, and children under the age of five being underserved (Rhodes et al., 2005). One possible reason for this disproportion is the various methods used by each school district and by each state to identify and service students with educational needs (Henderson et

al., 1993). While a federal definition has been provided, states vary greatly in their interpretation of the federal definition of LEP or EL students. State definitions of an "EL student" directly influence both the methods used for identification and the types of programs available (Rhodes et al., 2005). Differences in assessment methods and LEP criteria could result in an inconsistent understanding of who EL students are and what their needs may be (Henderson et al., 1993).

A second possible reason that may explain the disproportionality of EL students in special education programs is the variability between formal assessment practices, special education inclusion criteria, and referral procedures used to determine a student's need for such programing (Rhodes et al., 2005). As previously mentioned, many programs have begun using the RTII model while others still rely on the discrepancy model for identification (Kondisko, 2017; Salvia et al., 2013; Stevenson, 2015; Stokes, 2010). The use of measures that rely heavily on language seem likely to impact how an EL student performs. As such, EL students tend to underperform their English-speaking peers (Rhodes et al., 2005). Additionally, the assessments used to determine whether a student meets specified criteria for a disability can vary from district to district, resulting in different placements depending on a student's geographic location.

Using brief measures such as R-CBMs can be the simple solution needed to ensure that EL student needs are being addressed and assessed appropriately. Instruction that includes R-CBMs, such as the Curriculum-Based Decision Making (CBDM) model, allows teachers to regularly assess student performance, make instructional adaptations based on the data collected, and evaluate the effectiveness of instructional changes (Ergul, 2007). Current results support the use of R-CBMs as reliable measures of academic performance and progress monitoring for Spanish-speaking EL students.

Conclusion

R-CBMs provide a norm-referenced, quick, simple, and inexpensive way of reliably and validly collecting student data for most academic areas (Baker & Good, 1995; Codding et al., 2015; Deno, 1985; Deno, 2003; Ergul, 2007; Espin & Wallace, 2004; Fuchs, 2016; January & Ardoin, 2015; Salvia et al., 2013; Shapiro, 2011). They can be used as universal screeners to identify students who may be at risk for academic difficulties, as progress monitoring tools to assess student growth in a specific academic area, and as brief means of academic assessment that can accurately predict student performance on high-stakes assessments (Codding et al., 2015; Deno, 2003; Kim et al., 2016; Ramírez et al., 2007; Rhodes et al., 2005; Salvia et al., 2013; Shapiro, 2011). The beneficial utilities of R-CBMs have also been shown to extend to EL students (Baker & Good, 1995; Campbell et al., 2013; Ergul, 2007; Graves et al., 2005; Keller-Margulis & Mercer, 2014; Kondisko, 2017; Laija-Rodríguez et al., 2006; Ordóñez et al., 2002; Shapiro et al., 2006).

Current findings provide additional support for the use of R-CBMs with Spanish-speaking EL students. R-CBMs can help educators understand how EL students compare to their peers and can assist in determining if the student is making adequate yearly progress (Nese et al., 2011; Sandberg & Reschly, 2011). Along with other clinical assessment tools, such as language assessments and acculturation forms, R-CBMs can also be used as one component of the comprehensive educational evaluation when assessing EL students for a learning disability (Espin et al., 2012). Understanding an EL students' unique educational needs are necessary in ensuring the success of the student and promoting positive future outcomes.

References

Alyousef, H. S. (2006). Teaching reading comprehension to ESL/EFL learners. *Journal of Language and Learning, 5*, 63-73. http://www.readingmatrix.com/articles/alyousef/article.pdf

Ardoin, S. P., Roof, C. M., Klubnick, C., & Carfolite, J. (2008). Evaluating curriculum-based measurement from a behavioral assessment perspective. *The Behavior Analyst Today, 9*, 36. https://psycnet.apa.org/fulltext/2010-12932-006.pdf

Baker, D. L., Biancarosa, G., Park, B. J., Bousselot, T., Smith, J. L., Baker, S. K., Kame'enui, E. J., Alonzo, J., & Tindal, G. (2015). Validity of CBM measures of oral reading fluency and reading comprehension on high-stakes reading assessments in Grades 7 and 8. *Reading and Writing, 28*, 57-104. https://doi.org/10.1007/s11145-014-9505-4

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. https://eric.ed.gov/?id=ED372369

Bryington, A. A., Palmer, D. J., & Watkins, M. W. (2002). The estimation of interobserver agreement in behavioral assessment. *The Behavior Analyst Today, 3*, 323. http://dx.doi.org/10.1037/h0099978

Bujang, M. A. & Baharum, N. (2017). A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: A review. *The Journal of the School of Dental Sciences, 12*, 1-11. https://www.researchgate.net/publication/318788161_A_simplified_guide_to_determination_of_sample_size_requirements_for_estimating_the_value_of_intraclass_correlation_coefficient_A_review

Campbell, H., Espin, C. A., & McMaster, K. (2013). The technical adequacy of curriculum-based writing measures with English learners. *Reading and Writing, 26*, 431-452. https://doi.org/10.1007/s11145-012-9375-6

Chiappe, P., Siegel, L., & Wade-Woolley, L. (2002). Linguistic diversity and the development of reading skills: A longitudinal study. *Scientific Studies of Reading, 6*, 369–400. https://doi.org/10.1207/S1532799XSSR0604_04

Codding, R. S., Petscher, Y., & Truckenmiller, A. (2015). CBM reading, mathematics, and written expression at the secondary level: Examining latent composite relations among indices and unique predictions with a state achievement test. *Journal of Educational Psychology, 107*, 437. https://doi.org/10.1037/a0037520

Collier, C. (1998). *Acculturation: Implications for assessment, instruction, and intervention* (ED421871). National Association for Bilingual Education. ERIC. https://eric.ed.gov/?id=ED421871

Cummins, J. (1999). *BICS and CALP: Clarifying the distinction* (ED438551). ERIC. https://eric.ed.gov/?id=ED438551

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219-232. https://doi.org/10.1177/001440298505200303

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*, 184-192. https://doi.org/10.1177/00224669030370030801

Derwing, T. M. & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *Teachers of English to Speakers of Other Languages (TESOL) Quarterly, 39*, 379-397. http://doi.org/10.2307/3588486

Domínguez de Ramírez, R., & 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), 356-369. https://doi.org/10.1177/1534508411435721

EasyCBM (2016). *EasyCBM teacher deluxe user's manual: A BRT project*. University of Oregon. https://help.easycbm.com/wp-content/uploads/2016/01/easyCBM-Teacher-Deluxe-Users-Manual.pdf

Ergul, C. (2007). *Curriculum based decision making in an early literacy program* (Publication No. 3287935) [Unpublished doctoral dissertation, Arizona State University]. ProQuest Dissertations Publishing. https://search.proquest.com/openview/de26af20bcfdb76ec1663142d8e6cd36/1?pq-origsite=gscholar&cbl=18750&diss=y

Espin, C., McMaster, K. L., & Rose, S. (2012). *A measure of success: The influence of curriculum-based measurement on education*. University of Minnesota Press. https://doi.org/10.5749/j.cttttxf9

Espin, C., & Wallace, T. (2004). Descriptive analysis of curriculum-based measurement literature [Working document]. University of Minnesota Institute for Research on Progress Monitoring, Minneapolis, MN, United States.

Espin, C., Wallace, T., Lembke, E., Campbell, H., & Long, J. D. (2010). Creating a progress-monitoring system in reading for middle-school students: Tracking progress toward meeting high-stakes standards. *Learning Disabilities Research & Practice, 25*(2), 60-75. https://doi.org/10.1111/j.1540-5826.2010.00304.x

Every Student Succeeds Act, Pub. L. 114-95, codified as amended at 20 U.S.C.: Education § 1001 (2015). https://www.ed.gov/essa?src=rn

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. https://ecommons.luc.edu/luc_diss/663/

Fuchs, L. S. (2016). Curriculum-based measurement as the emerging alternative: Three decades later. *Learning Disabilities Research & Practice, 32*(1), 5-7. https://doi.org/10.1111/ldrp.12127

Gansle, K. A., VanDerHeyden, A. M., Noell, G. H., Resetar, J. L., & Williams, K. L. (2006). The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students. *School Psychology Review, 35*, 435. https://www.researchgate.net/publication/234729295_The_Technical_Adequacy_of_Curriculum-Based_and_Rating-Based_Measures_of_Written_Expression_for_Elementary_School_Students

Gersten, R., Baker, S. K., Shanahan, T., Linan-Thompson, S., Collins, P., & Scarcella, R. (2007). *Effective literacy and English language instruction for English learners in the elementary grades: IES practice guide*. What Works Clearinghouse. Princeton, NJ, United States. https://ies.ed.gov/ncee/wwc/PracticeGuide/6#tab-details

Geva, E., Yaghoub-Zadeh, Z., & Schuster, B. (2000). Part IV: Reading and foreign language learning: Understanding individual differences in word recognition skills of ESL children. *Annals of Dyslexia, 50,* 121–154. https://doi.org/10.1007/s11881-000-0020-8

Goldenberg, C. (2013). Unlocking the research on English learners: What we know - and don't yet know - about effective instruction (EJ1014021). *American Educator, 37*, 4. ERIC. https://files.eric.ed.gov/fulltext/EJ1014021.pdf

Graves, A. W., Plasencia-Peinado, J., Deno, S. L., & Johnson, J. R. (2005). Formatively evaluating the reading progress of first-grade English learners in multiple-language classrooms. *Remedial and Special Education, 26*, 215-225. https://doi.org/10.1177/07419325050260040401

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8,* 23. https://doi.org/10.20982/tqmp.08.1.p023

Hansen, L. (2006). Strategies for ELL success (EJ758192). *Science and Children, 43,* 22-25. ERIC. https://eric.ed.gov/?id=EJ758192

Healy, K. D. (2007). Word identification fluency and nonsense word fluency as predictors of reading fluency in first grade (EJ696631). ERIC. https://eric.ed.gov/?id=EJ696631

Henderson, A., Abbot, C., & Strang, W. (1993). Summary of the bilingual education state educational agency program survey of states' limited English proficient persons and available resources 1991–1992 (ED369292). Office of Bilingual Education and Minority Languages Affairs. ERIC. https://eric.ed.gov/?id=ED369292

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review, 34,* 507. https://www.researchgate.net/publication/238546006_Psychometrics_of_Direct_Observation

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABCs of CBM: A practical guide to curriculum-based measurement.* Guilford Publications.

Improving America's Schools Act, Pub. L. 103-382, 108 Stat. 3518, codified as amended at 20 U.S.C.: Education § 6301 (1994). https://uscode.house.gov/view.xhtml?path=/prelim@title20/chapter70&edition=prelim

January, S. A., & Ardoin, S. P. (2015). Technical adequacy and acceptability of curriculum-based measurement and the measures of academic progress. *Assessment for Effective Intervention, 41,* 3-15. https://doi.org/10.1177/1534508415579095

Jimerson, S., 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 (NAER Journal), 2*, 3-11. https://foi.org/10.7821/naer.2.1.3-11

Keller-Margulis, M. A., Clemens, N. H., Im, M. H., Kwok, O. M., & Booth, C. (2012). Curriculum-based measurement yearly growth rates: An examination of English language learners and native English speakers. *Learning and Individual Differences, 22*, 799-805. https://doi.org/10.1016/j.lindif.2012.07.005

Keller-Margulis, M. A., & Mercer, S. H. (2014). R-CBM in Spanish and in English: Differential relations depending on student reading performance. *Psychology in the Schools, 51*, 677-692. https://doi.org/10.1002/pits.21780

Kim, H. Y. (2013). Statistical notes for clinical researchers: Evaluation of measurement error 1 using intraclass correlation coefficients. *Restorative Dentistry & Endodontics, 38*(2), 98-102. https://doi.org/10.5395/rde.2013.38.2.98

Kim, J. S., Vanderwood, M. L., & Lee, C. Y. (2016). Predictive validity of curriculum-based measures for English learners at varying English proficiency levels. *Educational Assessment, 21*, 1-18. https://doi.org/10.1080/10627197.2015.1127750

Klingner, J. K., Hoover, J. J., & Baca, L. M. (2008). Why do English language learners struggle with reading?: Distinguishing language acquisition from learning disabilities. Corwin Press.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163. http://doi.org/10.1016/j.jcm.2016.02.012

Kondisko, J. E. (2017*). A preliminary investigation of racial bias in early writing curriculum-based measures* [Unpublished doctoral dissertation, Wilkes University] (ED576396). ERIC. https://eric.ed.gov/?id=ED576396

Lafrance, A., & Gottardo, A. (2005). A longitudinal study of phonological processing skills and reading in bilingual children. *Applied Psycholinguistics, 26,* 559–578. https://doi.org/10.1017/S0142716405050307

Laija-Rodríguez, W., Ochoa, S. H., & Parker, R. (2006). The crosslinguistic role of cognitive academic language proficiency on reading growth in Spanish and English. *Bilingual Research Journal, 30,* 87-106. https://doi.org/10.1080/15235882.2006.10162867

Lewandowski, L. J., Codding, R. S., Kleinmann, A. E., & Tucker, K. L. (2003). Assessment of reading rate in postsecondary students. *Journal of Psychoeducational Assessment, 21,* 134-144. https://doi.org/10.1177/073428290302100202

Massachusetts Department of Elementary and Secondary Education [DESE]. (2018). Massachusetts comprehensive assessment system: Test administrator's manual. Massachusetts Department of Elementary and Secondary Education. http://www.doe.mass.edu/mcas/testadmin/manual/TAM-CBT-g3-8elamath.pdf

McDermott, P. A. (1988). Agreement among diagnosticians or observers: Its importance and determination. *Professional School Psychology, 3,* 225-240. http://dx.doi.org/10.1037/h0090563

McFarland, J., Hussar, B., de Brey, C., Snyder, T., Wang, X., Wilkinson-Flicker, S., Gebrekristos, S., Zhang, J., Rathbun, A, Barmer, A., Mann, F. B., & Hinz, S. (2017). *The condition of education: 2017.* National Center for Education Statistics. https://nces.ed.gov/pubs2017/2017144.pdf

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46. http://dx.doi.org/10.1037/1082-989X.1.1.30

Mennuti, R. B., Christner, R. W., & Freeman, A. (2012). *Cognitive-behavioral interventions in educational settings: A handbook for practice* (2nd ed.). Routledge Taylor & Francis Group.

Microsoft (2018). *Test your document's readability* [computer software]. Microsoft. https://support.office.com/en-us/article/test-your-document-s-readability-85b4969e-e80a-4777-8dd3-f7fc3c8b3fd2

Miller, K. C., Bell, S. M., & McCallum, R. S. (2015). Using reading rate and comprehension CBM to predict high-stakes achievement. *Journal of Psychoeducational Assessment, 33*, 707-718. https://doi.org/10.1177/0734282915574028

Minnesota Department of Education (2018). *Statewide testing*. https://education.mn.gov/MDE/fam/tests/

Moore, L. M. (1997). An evaluation of the efficacy of curriculum based measurement reading measures in the assessment of Hispanic children [Unpublished doctoral dissertation]. Indiana University of Pennsylvania.

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. *The California School Psychologist, 14*, 11-21. https://doi.org/10.1007/BF03340947

National Association of School Psychologists. (2017). *Shortages in school psychology: Challenges to meeting the growing needs of U.S. students and schools* [Research summary]. https://www.nasponline.org/resources-and-publications/resources-and-podcasts/school-psychology/shortages-in-school-psychology-resource-guide

Nese, J. F., Park, B. J., Alonzo, J., & Tindal, G. (2011). Applied curriculum-based measurement as a predictor of high-stakes assessment: Implications for researchers and teachers. *The Elementary School Journal, 111*, 608-624. https://doi.org/608-624.10.1086/659034

New York City Department of Education. (2017). *Division of English language learners and student support: English language learner demographics report for the 2016-17 school year.* NYCED. https://infohub.nyced.org/docs/default-source/default-document-library/2016-17-demographic-report-v10_remediated.pdf

No Child Left Behind Act, Pub. L. 107-110, 20 U.S.C.: Education § 6319 (2002). https://files.eric.ed.gov/fulltext/ED556108.pdf

Ordóñez, C. L., Carlo, M. S., Snow, C. E., & McLaughlin, B. (2002). Depth and breadth of vocabulary in two languages: Which vocabulary skills transfer? *Journal of Educational Psychology, 94*, 719-728. https://doi.org/10.1037/0022-0663.94.4.719

Pearson, N. C. S. (2012). *AIMSweb technical manual.* Pearson Assessments. https://www.aimsweb.com/wp-content/uploads/aimsweb-technical-manual.pdf

Pennsylvania System of School Assessment (2019). Pennsylvania system of school assessment: Handbook for assessment coordinators. Pennsylvania Department of Education. https://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PSSA/PSSA%20Handbook%20for%20Assessment%20Coordinators.pdf

Quirk, M., & Beem, S. (2012). Examining the relations between reading fluency and reading comprehension for English language learners. *Psychology in the Schools, 49*, 539-553. https://doi.org/10.1002/pits.21616

Ramírez, D., Domínguez, R., & Shapiro, E. S. (2007). Cross-language relationship between Spanish and English oral reading fluency among Spanish-speaking English language learners in bilingual education classrooms. *Psychology in the Schools, 44*, 795-806. http://dx.doi.org/10.1002/pits.20266

Rhodes, R. L., Ochoa, S. H., & Ortiz, S. O. (2005). *Assessing culturally and linguistically diverse students: A practical guide.* Guilford Press.

Richardson, R. D., Hawken, L. S., & Kircher, J. (2012). Bias using maze to predict high-stakes test performance among Hispanic and Spanish-speaking students. *Assessment for Effective Intervention, 37,* 159-170. https://doi.org/10.1177/1534508411430320

Rejaie, R., & Ortega, A. (2003). *PALS: Peer-to-peer adaptive layered streaming* [Paper presentation]. Proceedings from the 13th International Workshop on Network and Operating Systems Support for Digital Audio and Video. Monterey, CA, United States. https://doi.org/10.1145/776322.776347

Salvia, J., Ysseldyke, J., & Bolt, S. (2013). *Assessment in special and inclusive education.* Wadsworth, Cengage Learning.

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. https://doi.org/10.1177/0741932510361260

Shapiro, E. S. (2011). *Academic skills problems (4th ed.): Direct assessment and intervention workbook.* Guilford Press.

Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment, 24,* 19-35. https://doi.org/10.1177%2F0734282905285237

Shinn, M. R. (1989). *Curriculum-based measurement: Assessing special children.* Guilford Press.

Shinn, M. M., & Shinn, M. R. (2002). *AIMSweb training workbook: Administration and scoring of reading curriculum-based measurement (R-CBM) for use in general outcome measurement*.

Edformation. https://www.researchgate.net/publication/267793797_AIMSweb_Training_Workbook_Administration_and_Scoring_of_Reading_Curriculum-Based_Measurement_R-CBM_for_Use_in_General_Outcome_Measurement

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428. https://doi.org/10.1037//0033-2909.86.2.420

Slavit, D., & Ernst-Slavit, G. (2007). Teaching mathematics and English to English language learners simultaneously. *Middle School Journal, 39*, 4-11. https://doi.org/10.1080/00940771.2007.11461618

Smith, K., & Wallin, J. (2011). *DIBELS Next assessment manual*. Dynamic Measurement Group. http://vd-p.d91.k12.id.us/TitleI_Resources/Reading_Progress_Monitoring/DAZE/Dibels%20DAZE%20Assessment%20Manual.pdf

Spinelli, C. G. (2008). Addressing the issue of cultural and linguistic diversity and assessment: Informal evaluation measures for English language learners. *Reading & Writing Quarterly, 24*, 101-118. https://doi.org/10.1080/10573560701753195

Stevenson, N. A. (2015). Predicting proficiency on statewide assessments: A comparison of curriculum-based measures in middle school. *The Journal of Educational Research, 108*, 492-503. https://doi.org/10.1080/00220671.2014.910161

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 (Publication No. 3433188) [Unpublished doctoral dissertation, Loyola

University Chicago]. ProQuest Dissertations Publishing. https://search.proquest.com/docview/823969098

Suen, H. K. (1988). Agreement, reliability, accuracy, and validity: Toward a clarification. *Behavioral Assessment, 10,* 343-366. https://psycnet.apa.org/record/1989-24559-001

Swanson, H. L., Sáez, L., & Gerber, M. (2004). Do phonological and executive processes in English learners at risk for reading disabilities in grade 1 predict performance in grade 2? *Learning Disabilities Research & Practice, 19,* 225–238. https://doi.org/10.1111/j.1540-5826.2004.00108.x

Turuk, M. C. (2008). The relevance and implications of Vygotsky's sociocultural theory in the second language classroom. *Arecls, 5,* 244-262. https://pdfs.semanticscholar.org/3987/5cacea3cc95ae54e504af6259ae64912adb0.pdf

U.S. Department of Education, National Center for Education Statistics (2012). *Schools and Staffing Survey (SASS): Public school data file* [Data set]. NCES. https://nces.ed.gov/surveys/sass/tables/sass1112_20170314001_s1s.asp

Vanderwood, M., Tung, C., & Hickey, R. (2014). Use of CBA/ CBM with culturally and linguistically diverse populations. Routledge, Taylor & Francis Group.

Walqui, A. (2006). Scaffolding instruction for English language learners: A conceptual framework. *International Journal of Bilingual Education and Bilingualism,* 9, 159-180. https://doi.org/10.1080/13670050608668639

Wayman, M. M., McMaster, K. L., Sáenz, L. M., & Watson, J. A. (2010). Using curriculum-based measurement to monitor secondary English language learners' responsiveness to peer-mediated reading instruction. *Reading & Writing Quarterly, 26,* 308-332. https://doi.org/10.1080/10573569.2010.500260

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. https://doi.org/10.1177/07419325050260040301

Woodcock, R. W. (2005). *Woodcock-Muñoz Language Survey–Revised*. Riverside Publishing.

Wright, J. (2013). *How to assess reading speed with CBM: Oral reading fluency passages*. National Professional Resources. http://www.jimwrightonline.com/mixed_files/lansing_IL/_Lansing_IL_Aug_2013/2_CBA_ORF_Directions.pdf

Yeo, S. (2011). Reliability generalization of curriculum-based measurement reading aloud: A meta-analytic review. *Exceptionality, 19,* 75-93. https://doi.org/10.1080/09362835.2011.562094

Appendix A

How to Score the Reading-CBM Passages *(Administration 1)*

Score each reading passage along with the audio recordings provided. Please use the correct score form provided on Qualtrics to score the Reading-CBM Passage. Your most important task is to determine the number of Words Read Correctly (WRC). Examiners put a slash (/) through incorrect words. You may listen to the audio as many times as needed. The following provides guidance on how to score the passages.

General Scoring Rules

WHAT IS A WORD READ CORRECTLY (WRC)?

- Correctly Pronounced Words within context
- Self-corrected Incorrect Words within 3 seconds

WHAT IS AN ERROR?

- Mispronunciation of the Word or Substitutions
- Omissions
- 3-Second Pauses or Struggles

WHAT IS NOT INCORRECT (NIETHER A WRC OR ERROR)?

- Repetitions
- Dialect Differences
- Insertions (Consider them Qualitative Errors)

Calculating and Reporting R-CBM Scores

Please count the total number of words read and then subtract the total number of Errors to obtain the Words Read Correctly (WRC). Place in scores in the ***WRC/Errors*** format. For example, if the student read 10 words in 1 minute but had 3 errors, the student's score would be *7/3*.

Procedures for Submitting Scored Reading-CBM Passages *(Qualtrics submission)*

Below are the instructions for submitting your Reading-CBM scored passages to the researcher. Please read the directions carefully. If any questions or concerns arise, please email the researcher directly at Amm7518@psu.edu.

Uploading File

All scored Reading-CBM files should be submitted directly into the Qualtrics survey. After each audio passage is presented, a there will be an opportunity to submit a screenshot or picture of your scored passage. All Reading-CBM files must be:

- Saved as a PDF, text document (i.e., DOC, DOCX, TXT, ODT), or graphic (i.e., JPG, PNG, GIF)
- Saved using the following format: (1) your personalized rater code and (2) the passage # scored format; such as "**rater code_passage #**"
    - ***Example*: 1A2B3C_passage 8**

Appendix B

How to Score the Reading-CBM Passages *(Administration 2)*

Score each reading passage along with the audio recordings provided. Please use the correct score form provided on Qualtrics to score the Reading-CBM Passage. Your most important task is to determine the number of Words Read Correctly (WRC). Examiners put a slash (/ OR) through incorrect words. You may listen to the audio as many times as needed. The following provides guidance on how to score the passages.

General Scoring Plans

WHAT IS A WORD READ CORRECTLY (WRC)?

- Correctly pronounced words within context
- Self-corrected incorrect words within about 3 seconds
- Words read with rolled "r" sounds
- Words read without ending sounds (i.e., -s, -ed, etc.) that DO NOT change the word's meaning (e.g., "sleeping" → "sleepin__")
- Words with "th" sound replaced with "d" sound

WHAT IS AN ERROR?

- Mispronunciation of the word or substitutions
- Omissions
- 3-second pauses or struggles
- Words read with inserted sounds that DO change the word's meaning (e.g., "tree" → "trees")

WHAT IS NEITHER A WRC NOR AN ERROR (i.e., NOT COUNTED IN TOTAL WORDS READ COUNT)?

- Repetitions
- Insertions (Consider them Qualitative Errors)
- Dialect Difference

Calculating and Reporting R-CBM Scores

Please count the total number of words read and then subtract the total number of Errors to obtain the Words Read Correctly (WRC). Place in scores in the ***WRC/Errors*** format. For example, if the student read 10 words in 1 minute but had 3 errors, the student's score would be *7/3.*

Procedures for Submitting Scored Reading-CBM Passages *(Qualtrics submission)*

Below are the instructions for submitting your Reading-CBM scored passages to the researcher. Please read the directions carefully. If any questions or concerns arise, please email the researcher directly at Amm7518@psu.edu.

Uploading File

All scored Reading-CBM files should be submitted directly into the Qualtrics survey. After each audio passage is presented, a there will be an opportunity to submit a screenshot or picture of your scored passage. All Reading-CBM files must be:

- Saved as a PDF, text document (i.e., DOC, DOCX, TXT, ODT), or graphic (i.e., JPG, PNG, GIF)
- Saved using the following format: (1) your personalized rater code and (2) the passage # scored format; such as "**rater code_passage #**"
    - ***Example*: 1A2B3C_passage 8**

VITA

ASHLEY MARINEZ
Amm7517@psu.edu

EDUCATION

May 2020	Ph.D., School Psychology, The Pennsylvania State University
May 2017	M.Ed., School Psychology, The Pennsylvania State University
December 2014	B.A., Psychology, Queens College

PROFESSIONAL EXPERIENCE

2019-2020	School Psychologist/ CSE Chairperson, *Schenectady City School District,* Schenectady, NY
2016-2019	Graduate Research Assistant, *Clearinghouse for Military Family Readiness at Penn State*, State College, PA
2015-2016	Graduate Research Assistant, *Educational Psychology, Counseling, and Special Education (EPCSE) Department,* State College, PA

PAPER PRESENTATIONS

Marinez, A. "Teacher and Student Perceptions of the Class Environment". Poster presentation at the annual National Association of School Psychologists (NASP) conference, February 14, 2018. *Chicago, IL.*

Marinez, A. "Teacher and Student Perceptions of the Class Environment". Poster presentation at the annual Association of School Psychologists of Pennsylvania (ASPP) conference, October 25, 2017. *State College, PA.*

HONORS AND AWARDS

2019-2020	Dorothy E. Kinley-Schumacher Scholarship in the College of Education
2015-2019	Robert W. Graham Endowed Graduate Fellowship
2012-2014	Psi Chi Honor Society
2012-2014	Golden Key International Honor Society
2012-2014	National Honor Society of Collegiate Scholar