Toward the Fair and Valid Use of Curriculum-Based Measurement for Students with Intensive Writing Needs and Linguistically Diverse Backgrounds

A Dissertation

SUBMITTED TO THE FACULTY OF THE UNIVERSITY OF MINNESOTA

BY

Seohyeon Choi

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Dr. Kristen L. McMaster

April, 2025

Acknowledgments

There are many people to whom I owe my deepest thanks, and words fall short of expressing just how grateful I am. First and foremost, I would like to thank my advisor, Dr. Kristen L. McMaster, for her guidance and unwavering support. Kristen, you have shown me what it means to pursue research with both intellectual passion and a warm heart—driven by a genuine desire to support students and teachers, while approaching problems with rigor and integrity. I could always trust and rely on you, and I am deeply grateful for all you have taught me, both as a scholar and as a person.

I am grateful to my committee members, Drs. Alisha Wackerle-Hollman, Nana Kim, and Hooi Ling Soh. Thank you for the time and insight you brought to this dissertation, and for the ways you challenged and supported me throughout this journey. I would also like to thank the University of Minnesota Graduate School for granting me the 2024-2025 doctoral dissertation fellowship, which has been instrumental in supporting my dissertation.

I would also like to thank Dr. Kristen L. McMaster again, along with Dr. Erica Lembke and the Early Writing Project team, for giving me the opportunity to analyze the data that formed the basis of this dissertation. During my PhD journey, I also had the great fortune of being part of the Inference Galaxy team and working with wonderful colleagues and mentors. It has been a source of guidance, inspiration, and truly meaningful memories. I feel incredibly lucky to have learned from Drs. Panayiota Kendeou, Nidhi Kohli, and HyeJin Hwang—thank you for your mentorship and encouragement.

I am deeply thankful to Dr. Dong-il Kim, my master's advisor, who first sparked my interest in research—especially with a focus on better serving students whose needs are often overlooked or underrepresented. Thank you for showing me how to stay motivated by the hope

of making the field better. You also taught me the power of moving forward together, and gave me the courage to take an uncertain path.

Last, I am forever grateful to my family. If I have done anything right—or ever will—it is because of you. Through your lives, you have shown me what it means to live with honesty, integrity, and deep care for others. Because of you, I know it's okay to stumble or fall—there is always a place where I belong and can begin again. Thank you for being the model of the person I hope to become. Your unconditional love and quiet yet unwavering support has carried me through more than you know.

Abstract

Curriculum-based measurement (CBM) is a valuable assessment method for students with intensive learning needs, including in writing. Despite research on CBMs in writing, insufficient attention has been given to linguistic diversity, especially among young or beginning writers, leaving uncertainty about whether current assessment practices with writing CBMs provide reliable and valid information on multilingual students' early writing development in English. The purpose of this study was to evaluate the measurement invariance of Word Dictation, a specific CBM writing task designed to assess English transcription (spelling and handwriting) skills at the word level, across two groups: multilingual and English-monolingual students, both with intensive needs in writing. This study used data obtained from a large multi-site, multi-year randomized control trial, which investigated the effects of a professional development program designed to support elementary teachers in implementing data-based instruction for struggling early writers. This study analyzed Word Dictation responses from 349 students, including 67 multilingual and 282 English-monolingual students, primarily in Grades 1–3. In evaluating measurement invariance at both item and assessment levels, I used different scoring metrics, specifically words spelled correctly and correct letter sequences, and employed a range of analytical methods, including those based on item response theory and classical test theory frameworks. Within the current study context, I identified a few items as potentially displaying differential item functioning, but the magnitude was overall small. The direction of the detected differential item functioning was not systematic, with some items favoring the multilingual student group and others favoring the English-monolingual student group. When comparing the scoring metrics, some differential item functioning observed in words spelled correctly was mitigated when using correct letter sequences. At the assessment level, Word Dictation did not

function differently across the two student groups. These findings build upon existing knowledge of the technical adequacy of Word Dictation, incorporating varying scoring metrics, and further expand by providing evidence of validity and fairness for multilingual students learning to write in English. The discussion also highlights significant limitations of this study. These include grouping of students with diverse native languages into a single category of multilingual students, analyzing subsets of items rather than the entire set, and limitations in the specific analytic approach, such as limitations in selecting anchor items for analyses using correct letters sequences. With these limitations in mind, I discuss future research directions and implications for educators using Word Dictation to better serve linguistically diverse students requiring intensive support in developing their writing skills in English.

Table of Contents

List of Tables

List of Figures

Definitions of Key Terms

Evidence of Validity – The degree to which theoretical and empirical support is provided for the interpretations or inferences drawn from assessment scores, ensuring they are appropriate given their intended uses.

- *Evidence of Structural Validity* – Source of validity evidence related to whether the internal structure of the assessment accurately reflects the internal structure of the construct being measured.
- *Evidence of Fairness* – Aspect of validity evidence showing whether the assessment yields valid inferences for its intended purpose for *all* individuals within the intended populations, with the internal structure remaining consistent across subgroups.

Measurement Invariance – One method for examining evidence of structural validity and fairness. Ensures item scores maintain consistent meaning across subgroups within the intended population.

- *Differential Item Functioning (DIF)* – Measurement invariance violated at the item level. Indicates certain items function differently for students with similar ability levels solely because of group membership.
- *Differential Test Functioning (DTF)* – Measurement invariance violated at the assessment level. Suggests total scores of the assessment as a whole may not hold equal meaning across subgroups within the intended population.

Rasch Model – A specific type of Item Response Theory (IRT) model designed for dichotomous items (two possible responses). Estimates a person's ability and the difficulty of each item and

expresses the probability of a correct response as a logistic function of the difference between the person's ability and the item's difficulty.

Partial Credit Model – An extension of the Rasch model for polytomous items (more than two response categories). Divides each item into ordered thresholds or steps, with each threshold or step representing a transition between response categories. Estimates a person's ability and the difficulty of each step within an item.

Correct Letter Sequences (CLS) – A specific type of scoring method for curriculum-based measurements in writing. Scores two adjacent, correctly placed letters within a word as a correct letter sequence.

Transcription – A process of transforming language representations in working memory into written or typed orthographic symbols using a pen, pencil, or keyboard.

- *Spelling* – A component skill of transcription that involves the ability to encode sounds into written words following the orthographic system of a language.
- *Handwriting* – Another component skill of transcription that relates to the ability to form and write letters accurately and with speed. Involves orthographic coding along with the motor system.

Chapter 1: Introduction

Writing is critical for students' overall literacy development (Biancarosa & Snow, 2004; Fitzgerald & Shanahan, 2000) and their long-term success in school and beyond (Graham et al., 2023). However, developing writing skills is not an easy task, requiring the integration and coordination of various literacy-, language-, and cognition-related skills (Kim & Graham, 2022; McCutchen, 2006). This complexity, coupled with the often inadequate support for writing in many schools (Kent et al., 2017), has led to a nontrivial proportion of children in the United States (U.S.) facing challenges in writing. For instance, approximately 75% of students in U.S. schools have not reached proficiency levels in writing by eighth grade (National Center for Education Statistics, 2003, 2021).

The challenge in developing English writing skills is particularly pronounced among some groups of students, including students identified with different types of disabilities (e.g., Graham et al., 2016, 2017, 2020; Grigorenko et al., 2020; Pennington & Delano, 2012), multilingual students, and those with intersecting identities, being both multilingual and with disabilities. For multilingual students with or without disabilities, the process of learning to write in a non-native language can be particularly challenging compared to their monolingual peers. Data from the 2011 National Assessment of Educational Progress (NAEP) in writing indicates that, as a group, multilingual students have demonstrated lower achievement compared to English-monolingual students, with only 1% reaching proficient levels by the eighth grade (National Center for Education Statistics, 2012).

In response to the need to support students' writing development, many researchers have invested significant efforts, focusing on identifying various research- or evidence-based writing

interventions (e.g., see Datchuk & Kubina, 2012; Graham et al., 2012; McMaster et al., 2018 for literature syntheses of writing interventions for broad student populations with or without writing difficulties). However, relatively less attention has been given to *assessment* compared to intervention. The current dissertation focuses on assessment practices for English writing, specifically targeting students identified with severe writing challenges.

Among the various assessments used in schools, the focus of the current dissertation is on one assessment framework called curriculum-based measurement (CBM; Deno, 1985), which is considered particularly valuable for identifying students who are struggling with academic learning, monitoring their progress, and guiding teachers' instructional decision making. Just as for CBMs in other academic areas such as reading and math, substantial evidence has accumulated to support the psychometric rigor of writing CBMs (McMaster & Espin, 2007; McMaster et al., 2011b; Romig et al., 2017, 2021; Shin & McMaster, 2019).

However, it remains less clear to what extent the currently available writing CBMs provide reliable and valid data for multilingual students learning to write in English. Previous studies investigating the psychometric rigor of English writing CBMs have predominantly focused on English-monolingual students, inadequately representing multilingual student populations (e.g., Keller-Margulis et al., 2016; Sandberg & Reschly, 2011). Findings from these studies may not necessarily be generalizable to multilingual student populations, and a separate empirical evaluation is required. This investigation is particularly crucial as many multilingual students are anticipated to possess unique needs within the context of English writing assessments (elaborated further in the "Understanding Students' Writing Development" section in this introduction); when there are specific subgroups expected to have differing needs within

assessment contexts, conducting separate validity and reliability analyses for specific subgroups is recommended (American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME], 2014).

Fortunately, some researchers have taken steps to assess the technical evidence of English writing CBMs specifically within multilingual student populations. These efforts have been systematically summarized by Smith et al. (2023), as well as in my own systematic literature review in Chapter 2 of this dissertation. The simple conclusion inferred from these reviews may be that the existing reliability and validity evidence for various English writing CBMs used for multilingual students falls within an acceptable range, especially for using the measures at a specific point in time, such as for screening purposes. Still, an important question remains unanswered.

The reviewed studies identified that, for multilingual students, scores obtained from English writing CBMs generally show satisfactory correlations with scores from other writing assessments, including standardized, norm-referenced measures (Campbell, 2010; Campbell et al., 2013; Espin et al., 2008; Keller-Margulis et al., 2016; Landis, 2019; Smith & Lembke, 2021, 2022). These studies addressed an important source of validity evidence, namely criterion-related validity. However, what is concerning is that no study evaluated whether the measure actually assesses students' latent ability (e.g., English writing skills) *in an invariant manner* across linguistically diverse student populations. Such an investigation is imperative for ensuring structural validity (i.e., that the internal structure of the measure aligns with the underlying construct) and fairness (i.e., that the measure functions equivalently for individuals within the intended population without favoring or disfavoring specific subgroups) of the assessments

(AERA/APA/NCME, 2014; Jonson & Geisinger, 2022). The current dissertation aims to address this gap.

The purpose of this dissertation is to evaluate the measurement invariance of Word Dictation, a type of CBM designed to assess English transcription skills (spelling and handwriting) at the word level, especially suited for young elementary students or those in the beginning stage of developing English writing skills. To achieve this goal, I examine whether or not items within Word Dictation, as well as the assessment as a whole, scored using different widely-used scoring metrics (words spelled correctly [WSC] and correct letters sequences [CLS]), function differently for multilingual and English-monolingual student groups, both identified as having intensive needs in early writing in English.

Together with existing literature, I anticipate that this study could contribute to establishing effective assessment practices using English writing CBMs for linguistically diverse students with intensive writing needs, which is essential for effectively supporting their writing development.

Curriculum-Based Measurements in Writing

Researchers initially developed CBM to offer special education teachers a reliable data source to monitor students' progress toward their long-term individualized education goals within a specific academic domain, serving as an alternative to informal observations or standardized tests (Deno, 1985). To serve as an effective progress monitoring tool, CBM was developed with specific attributes (Deno, 1985; Hosp et al., 2016) that include: (a) alignment with the curriculum, ensuring that students are tested on what they are being taught; (b) technical adequacy, ensuring evidence of reliability and validity of the measures (for research syntheses of

technical evidence for CBMs across various academic domains, see Foegen et al., 2007, McMaster & Espin, 2007, Reschly et al., 2009, Romig et al., 2017, Romig et al., 2021, Shin & McMaster, 2019, Wayman et al., 2007, Yeo, 2010); (c) standardized procedures for administration and scoring; (d) performance sampling, counting the correct and incorrect student behaviors on clearly defined tasks (e.g., number of correct words written in response to a story starter) within a set time interval (typically 3, 5, 7, or 10 minutes); (e) inclusion of decision rules to interpret student performance and progress; (f) emphasis on repeated measurement over time to determine progress rates and performance levels; and (g) efficiency. By virtue of these features, CBM sets itself apart from many other assessments and is conceptualized as a *general outcome measure*, signifying that it offers an indicator of the student's overall performance and progress within a specific academic domain (Fuchs & Deno, 1991). Furthermore, researchers have demonstrated the instructional utility of CBMs, given that educators' use of CBM for monitoring student progress and making instructional decisions has improved outcomes of students with or without learning difficulties (e.g., Jung et al., 2018; Stecker et al., 2005). Drawing from this extensive body of research on CBM, many schools, especially those implementing tiered systems of support, have presently incorporated CBM into their screening and progress monitoring (Burns et al., 2016; Clemens et al., 2016; Fuchs & Vaughn, 2012).

Using the CBM framework is beneficial in any academic domain, but the specific focus of this dissertation is on CBMs in writing. Writing CBMs, typically consisting of a set of forms with equivalent difficulty, encompass various tasks (McMaster & Espin, 2007). These writing CBM tasks vary based on several characteristics, such as the specific level of writing skills they are intended to assess (e.g., word-, sentence-, or passage-level writing), the presence of prompts

(e.g., word dictation task, which involves writing down a set of words and does not require prompts, while others involve writing prompts), and the type or genres of prompts (e.g., picture prompts, picture-word prompts, story starters, expository prompts).

Just like CBMs in other academic domains, English writing CBMs are backed by a solid body of literature showing their technical evidence. Technical evidence of CBM can be categorized into various stages, as suggested by Fuchs (2004): Stage 1, which involves assessing the reliability and validity evidence of static scores (required for using the measures for screening); Stage 2, which focuses on investigating the reliability, validity, and sensitivity of slopes over time (required for using the measure for progress monitoring); and Stage 3, which centers on determining instructional utility. For English writing CBMs, substantial evidence has been established for Stage 1 (for reviews and syntheses, see McMaster & Espin, 2007; McMaster et al., 2011a; Romig et al., 2017, 2021), considerable but less than Stage 1 evidence for Stage 2 (e.g., Choi et al., 2023; McMaster et al., 2011b; Romig & Olsen, 2021), and limited but growing evidence for Stage 3 (e.g., Choi et al., 2024; Jung et al., 2018; McMaster et al., 2020; Stecker et al., 2005).

One noteworthy feature of writing CBMs is that students' responses can be assessed using a variety of scoring metrics. The commonly used and researched scoring procedures include words written (WW; Deno et al., 1980), WSC (Deno et al., 1980), correct word sequences (CWS; Videen et al., 1982), CLS (Deno et al., 1980) and correct minus incorrect word sequences (CIWS; Espin et al., 2000), each of which is described further in the literature review of this dissertation. When understanding the existing technical evidence of writing CBMs, it is advantageous to consider which specific scoring methods were examined. Previous research

syntheses have documented nontrivial variations in criterion validity coefficients depending on specific scoring metrics adopted (Romig et al., 2017, 2021).

Variability in technical evidence depending on scoring metrics also applies to the assessments for multilingual students, although the available studies are limited in volume and vary in the grade levels of students, possibly contributing to this variation. For example, Espin and colleagues (2008) found that for tenth-grade multilingual students, CIWS best predicted their state test performance (Minnesota Basic Standards Test). Keller-Margulis and colleagues (2016) corroborated this finding with fourth-grade multilingual students by using the State of Texas Assessments of Academic Readiness (STAAR) Writing. Conversely, Smith and Lembke (2021) observed that CIWS was not a good predictor of the performance of multilingual students in grades 1–3 on the Assessing Comprehension and Communication in English State-to-State (ACCESS) English Language Proficiency test.

Understanding Students' Writing Development

Understanding how students develop writing skills is crucial for effective assessment practices. By knowing the theoretical frameworks of writing development and identifying the developmental stages students are in, educators can prioritize assessment tools that target specific skills or proficiency levels that are within the students' reach, which is particularly vital when working with struggling beginning writers. For instance, as will be described below, beginning writers invest significant working memory resources in transcription (spelling and handwriting), making these skills essential components of early writing assessment. Furthermore, aside from comprehending general writing development, understanding potential differences in writing development between multilingual students and English-monolingual

students and how such differences may manifest in the context of English writing assessments is important. This knowledge can help teachers gain deeper insights into the needs and strengths of multilingual students based on their writing products.

Fundamental Skills Underlying Writing

Researchers have proposed various theoretical frameworks to explain students' writing development, which include the Simple View of Writing (Berninger & Amtmann, 2003), the Not-So-Simple View of Writing (Berninger & Winn, 2006), and the Direct and Indirect Effects Model of Writing (Kim & Schatschneider, 2017). Despite the variations in focus, specificity, and structure across the frameworks, several skills are consistently highlighted as particularly important for early writing development.

Key to early writing development is the acquisition of *transcription* and *text generation* skills. First, transcription refers to the process of transforming language representations in working memory into written or typed orthographic symbols using a pen, pencil, or keyboard (Berninger, 1999). Researchers have conceptualized transcription as involving two distinct skills, (a) handwriting and (b) spelling, both of which have been found to be strongly correlated with overall writing ability (e.g., Ahmed & Wagner, 2020; Kent & Wanzek, 2016). Interventions targeting these skills have also demonstrated positive effects on overall improvements in writing performance (e.g., Graham et al., 2012).

Handwriting encompasses the integration of orthographic coding–the ability to encode a printed word in memory and subsequently access the entire word pattern, a single letter, or letter clusters within that representation–, along with components of the motor system responsible for executing the process of translating those words into print (Berninger & Rutberg, 1992). Spelling

also entails orthographic coding, alongside phonological coding–the ability to analyze and synthesize phonemes in words. To be able to spell, students need to draw upon several key emergent literacy skills, including alphabetic knowledge, phonological awareness, print knowledge, name writing, and letter writing (Ehri, 2000; Lonigan et al., 2008; Puranik et al., 2011). As students automatize these skills, they can more efficiently allocate limited working memory resources to other processes involved in advanced writing (McCutchen, 1996). Word Dictation, the focused writing measure of this study, is designed to assess transcription skills in English, encompassing both spelling and handwriting.

Text generation encompasses two aspects: (a) idea generation or ideation, involving generating ideas using language within working memory, and (b) translation, involving converting these ideas into language representations (Berninger, 1999). Studies have shown positive correlations between text generation and overall writing (Hooper et al., 2011; Juel et al., 1986; Kim et al., 2011, 2015), even when controlling for other component skills like spelling (Juel et al., 1986; Kim et al., 2015). Oral language skills (e.g., vocabulary, morphosyntactic, syntactic knowledge) are often used as proxies for text generation (Graham & Eslami, 2022), although text generation also draws upon background and content knowledge (Kim et al., 2018; Kim & Schatschneider, 2017).

As students advance to middle and high school, the production of quality written compositions becomes important. During this period, domain-general cognitive skills (e.g., working memory, attention, linguistic awareness), metacognitive skills (e.g., self-regulation, planning, strategy use), and higher-order skills (e.g., inference-making, perspective-taking, self-

monitoring) (Hooper et al., 2011; Kim & Schatschneider, 2017) become more important than before.

It is important to note that the descriptions provided thus far are mainly derived from literature pertaining to writing *in English*. The trajectory and characteristics of writing development may vary depending on specific languages. Languages vary in different aspects (e.g., phonological structures, consistency of phoneme-grapheme correspondences, writing system; Kim, 2011), and such variations influence the trajectory of writing development (e.g., McBride-Chang et al., 2005).

For instance, in alphabetic writing systems like English, Spanish, and Korean, phonological awareness and letter knowledge play a particularly significant role in spelling development (Caravolas et al., 2001; Furnes & Samuelsson, 2009). On the other hand, in languages like Chinese that use a morpho-syllabic writing system, spelling demands a broader range of cognitive skills than in alphabetic languages, as it involves converting spoken language into visual configurations that may appear arbitrary and lack clear phonological clues (Ye et al., 2022). Furthermore, even among languages that adopt the alphabetic writing system, there are variations in writing development. For example, Kim (2011) identified that syllable awareness plays a particularly significant role in Korean spelling development, whereas in English spelling, phoneme awareness holds more prominence (Hulme et al., 2002). Due to these language-specific variations in writing, the presumptive transfer of assessment measures from one language to another should be avoided. For instance, Winkes and Schaller (2022) highlighted that the optimal conditions established for administering English writing CBMs may not be directly applicable to writing CBMs in German.

Potential Influence of Cross-Linguistic Transfer in Multilingual Students' L2 Writing

Compared to the extensive efforts within the research community to understand how early writing skills develop in general (Berninger & Amtmann, 2003; Berninger & Winn, 2006; Kim, 2020; Kim & Schatschneider, 2017), there remains a dearth of knowledge about early writing development in L2–language(s) other than the first language (L1)–among multilingual students (Fitzgerald et al., 2014; Graham & Eslami, 2020; Lesaux et al., 2006; Riazi et al., 2018; Williams & Lowrance-Faulhaber, 2018; Wolters & Kim, 2024). Some researchers (Graham & Eslami, 2020, 2022) have recently used the existing theoretical frameworks, particularly the Simple View of Writing (Berninger & Amtmann, 2003), to offer insights into multilingual students' L2 writing development. Overall, their findings suggest that the components of this framework, particularly transcription and text generation, do significantly explain the variation in English writing among English L2 learners, across countries and grade levels (Graham & Eslami, 2020, 2022). However, ongoing research would be beneficial, as literature highlights how the languages within a student's linguistic repertoire interact and collectively influence the development of multilingual writing skills (Lanauze & Snow, 1989; Schnoor & Usanova, 2023; Usanova & Schnoor, 2021). Therefore, neither L1 nor L2 theories alone can fully capture the complexities of the multilingual experience (Grosjean, 2010; Valdés & Figueroa, 1994).

Despite the lack of specific theoretical frameworks, many linguistic studies highlight at least one crucial concept that is relevant in understanding young multilingual students' L2 writing: the potential influence of cross-linguistic transfer (Chung et al., 2019; Cummins, 1979; Fitzgerald et al., 2014; MacWhinney, 2005). Various frameworks explain this phenomenon, including the linguistic interdependence hypothesis (Coady, 1997; Cummins, 1979; Verhoeven,

1994), unified model (MacWhinney, 2005), contrastive analysis (Connor, 1996; Lado, 1957), transfer facilitation model (Koda, 2008), common underlying cognitive process (Geva & Ryan, 1993), and the interactive transfer model (Chung et al., 2018). Essentially, the idea of cross-linguistic transfer suggests that L1 influences L2 language and literacy, and vice versa.

When multilingual students write in English, they often use their linguistic knowledge in both their native language and English strategically and bidirectionally (see Williams & Lowrance-Faulhaber, 2018 for a literature review). When applying the contrastive-typological framework (Lado, 1957), cross-linguistic transfer in writing may occur more frequently when there is greater overlap in the linguistic features (especially focusing on structural aspects within this framework) between L1 and L2 (Chung et al., 2019; Melby-Lervåg & Lervåg, 2011). Also, applying the unified theory (MacWhinney, 2005), cross-linguistic transfer is more likely to occur when a psychological unit in English triggers a relevant unit in the student's L1.

Taking Spanish-English bilingual students as an example, students' knowledge of Spanish can play a role in their English writing in several ways. Spanish and English both use Roman alphabet scripts, featuring significant overlap in graphemes and phonemes, as well as shared etymological roots in many words (Chung et al., 2019; Linan-Thompson & Meline, 2022). As such, Spanish-English bilingual students who possess developed language and literacy skills in Spanish are likely to source their Spanish knowledge to varying extents when they write in English, giving them some advantages in English writing as compared to multilingual students whose L1 uses non-Latin scripts (e.g., Korean, Arabic, Chinese, Russian).

However, Spanish graphophonemic knowledge does not always facilitate English writing; sometimes, it can instead lead to English spelling errors. This is because not all English

letters are represented in Spanish, and most English words cannot be spelled correctly using Spanish graphophoneme rules (Wolters & Kim, 2024). Some common English spelling errors indicating the influence of Spanish include: writing *pleasing* for *playing, slip* for *sleep,* and *da* for *the* (August & Shanahan, 2006; Linan-Thompson & Meline, 2022). Wolters and Kim (2024), who examined English and Spanish writing among Spanish-English bilingual students in grades 1–3 enrolled in dual immersion programs, identified consistent patterns of English spelling errors indicating cross-linguistic transfer. These patterns include long-e replacement (e.g., writing *hi* for *he*), interdentals replaced with dentals (e.g., writing *de* for *they*), and short-i replacement (e.g., writing *et* for *it*), among others.

Evaluating Measurement Invariance of Word Dictation

Evidence for Structural Validity and Fairness, and Measurement Invariance Evaluation

Measuring something essentially entails observing a particular characteristic and then representing the results of that observation using numerical values or symbols. For these measurements to be useful, it is fundamental to have evidence, ideally based on a cumulative body of research, showing that the measurement results provide reliable and valid data for the target populations. Specifically, reliability, validity, and fairness are considered central psychometric properties for educational assessment (AERA/APA/NCME, 2014). The main focus of the current dissertation is on validity and fairness evidence. However, it is important to note that the concepts of reliability, validity, and fairness are not entirely distinct but rather interconnected; fairness is an essential component of validity, aiming to prevent sources of construct-irrelevant variance, and high reliability is a prerequisite for validity (Jonson & Geisinger, 2022).

Validity refers to the extent to which theory and empirical evidence support the interpretations or inferences drawn from assessment scores for their intended uses (Messick, 1989). To establish validity, various sources of evidence should be considered, including those related to test content, response processes, relation with other variables, internal structure, and consequences of testing (AERA/APA/NCME, 2014). Specifically, evidence related to the internal structure, often termed structural validity, assesses whether the internal structure of the assessment reflects the internal structure of the underlying construct being measured (AERA/APA/NCME, 2014; Messick, 1995). Of important note, the internal structure should remain consistent across different subgroups within the intended populations unless the theory suggests distinct internal structures. Fairness in testing, being closely tied to structural validity, refers to whether the assessment yields valid inferences for the intended purpose for *all* individuals within the intended populations, with the measurement model remaining consistent across subgroups within the population (AERA/APA/NCME, 2014; Jonson & Geisinger, 2022; Zieky, 2016).

One method for examining evidence for structural validity and fairness is assessing the measurement invariance (i.e., measurement equivalence) of the assessment tool. Measurement invariance pertains to whether scores obtained from indicators (e.g., items, scales) of underlying constructs maintain invariant meaning across various conditions (Meredith, 1993; Vandenberg & Lance, 2000). Typically, such an evaluation involves comparing the scores of two subgroups within the intended population, with group membership determined by factors such as race/ethnicity, sex/gender, geographic region, or linguistic status that are irrelevant to the construct being measured in a given context. Without establishing measurement invariance, it

cannot be ensured that the assessment scores have the equivalent meaning across the subgroups (Alatli, 2000; Drasgow, 1984; Vandenberg & Lance, 2000).

This evaluation can be conducted at the item level, scale level (if there are subscales within the assessment), or at the entire assessment level. If measurement invariance is not upheld *at the item level*, often referred to as differential item functioning (DIF; Holland & Wainer, 2012), it suggests that certain items function differently for examinees with equivalent ability levels, solely because of the group membership (Angoff, 1993; van de Vijver & Leung, 2010). That is, one group may receive lower scores on the items or be less likely to answer the items correctly compared to another group matched in ability level. The detection of DIF implies that such discrepancy is *systematic,* affecting all or nearly all individuals in a similar, predictable manner (Osterlind & Everson, 2009). However, at the same time, it is essential to note that DIF is a group phenomenon (Osterlind & Everson, 2009); hence, making inferences about an individual's performance on the items is not valid.

Measurement invariance can also be compromised at the assessment level, often termed differential test functioning (DTF). Under DTF, the total scores of the assessment as a whole may not hold equal meaning across groups. Of note, DIF does not necessarily lead to DTF. In other words, despite the presence of DIF for some items, it is possible that the assessment-level scores have equivalent meaning across the groups. This situation can happen when the effects of DIF across items cancel each other out, with some items favoring one group while others favor the other group (Rodríguez-Casallas et al., 2020).

DIF is not synonymous with item *bias* (Jonson & Geisinger, 2022; Penfield & Lam, 2000; Zieky, 1993), and likewise, DTF is not synonymous with test *bias*. Observing DIF simply

indicates a mathematically significant discrepancy in the item responses between the groups. To establish bias, further in-depth investigation is required to determine the underlying reasons for the detected DIF, often through content experts' item reviews. Item bias can be confirmed only when a systematic association between the given item and a construct other than the one intended for assessment is identified. For example, as illustrated in Goodrich et al. (2019), items in a Spanish-language vocabulary assessment may exhibit DIF among Spanish-speaking children from various U.S. regions due to dialect variations; however, if dialect variation aligns with the assessment's intended purpose, the presence of DIF would not necessarily indicate bias.

DIF is indeed a very complex phenomenon, requiring a comprehensive evaluation and careful interpretation. It is not unusual for researchers to observe items being statistically flagged as exhibiting DIF, although, upon expert content review, the items appear to validly measure the construct without favoring or disadvantaging specific groups (Embreston & Reise, 2000). Also, even with in-depth investigations into underlying reasons, it often remains unclear whether the observed differences stem from genuine disparities between the groups in the intended construct or from biases, and quite frequently, it is a combination of both of them (AERA/APA/NCME, 2014; Embreston & Reise, 2000).

Importance of Embracing Various Approaches for Assessing Differential Item Functioning

Researchers have developed a wide range of statistical methods to detect DIF and quantify its magnitude. Different detection approaches can be broadly categorized into traditional methods that rely on observed scores and others that are based on latent traits. Observed score-based methods include, but are not limited to, logistic regression (Swaminathan & Rogers, 1990) and Mantel-Haenszel (Holland & Thayer, 1988). Approaches based on unobserved scores or

latent variables include multi-group confirmatory factor analysis (CFA) and techniques using item response theory (IRT) modeling like the likelihood ratio test, Raju's area method (Raju, 1988, 1990), and their extensions (e.g., Differential Functioning of Items and Tests framework [DFIT], Raju et al., 1995). These latent trait-based approaches are often preferred to logistic regression and Mantel-Haenszel (Millsap & Everson, 1993; Pendergast et al., 2017).

While diversity in analytic approaches is advantageous, it can often complicate the interpretation of DIF, as decisions about the presence or absence of DIF can vary depending on which methods are used. Furthermore, even within the same approach, DIF results can be influenced by numerous factors related to the study design. For instance, when adopting the IRT-based likelihood ratio test, the sample size (both respondent sample and assessment length), the proportion of true DIF items, and the balance or imbalance of DIF items can affect the performance of the likelihood ratio test, thereby impacting the decisions regarding the presence or absence of DIF (Wang et al., 2022).

Recognizing this complexity of DIF and possible variation in DIF decisions depending on analytic strategies employed, in the current study, I use multiple methods. My primary method is the IRT-based likelihood ratio test, chosen for its unique advantages such as the availability of a wide range of useful item-level metrics, model flexibility, ability to handle missing data effectively, and rigorous detection of DIF by using DIF-free items as anchors (e.g., Banks, 2015; Tay et al., 2015). However, my analysis also incorporates logistic regression and the Mantel-Haenszel method, which are rooted in the classical test theory (CTT) model and use observed scores, as well as Raju's area method and the DFIT framework, which are alternative IRT-based approaches (see the Method section for details).

Importance of Evaluating Measurement Invariance of Word Dictation

As previously mentioned, CBM is a well-supported assessment approach with a robust foundation in research supporting its reliability and validity, which also applies specifically to the domain of writing. However, existing research on English writing CBMs primarily centers on English-monolingual students (Keller-Margulis et al., 2016; Smith & Lembke, 2022), providing limited insight into the technical evidence of these measures for multilingual students learning to write in English. While some studies have started to address this gap by focusing on specific aspects like criterion-related validity (Smith et al., 2023; also see my own systematic review findings in Chapter 2), there is no empirical study evaluating measurement invariance of English writing CBMs across multilingual and English-monolingual students, especially those identified as having intensive early writing needs. Assessing measurement invariance and thus providing validity evidence for students from various linguistic backgrounds is a crucial part of Stage 1 research as identified by Fuchs (2004). Indeed, it is not unusual for researchers to detect DIF items in L2 literacy assessments for multilingual students (e.g., Farrington et al., 2015; Goodrich et al., 2019; Koo et al., 2014). In Word Dictation, it is also possible that certain items function differently depending on the student's language status; if that is the case, such observations should be considered carefully when using Word Dictation with students.

DIF in Word Dictation items, if identified, could stem from various factors. One such factor is the influence of cross-linguistic transfer. Cross-linguistic transfer, if it occurs, might affect how multilingual and English-monolingual students respond to Word Dictation items differently. For instance, suppose that Spanish-English bilingual students identified as having severe difficulties in English writing possess stronger language or writing skills in Spanish.

Additionally, suppose that some Word Dictation items involve letters with shared grapheme-phoneme correspondences in Spanish and English, or cognates that are similar in both languages. In this scenario, these students could leverage their knowledge of Spanish spelling rules to respond to those Word Dictation items (Wolters & Kim, 2024). Consequently, these students may have a higher probability of answering the given items correctly compared to English-monolingual students (who were also identified as having intensive writing needs) with equivalent levels of English transcription skills, thereby indicating DIF items favoring the Spanish-English bilingual student group.

Conversely, some Word Dictation items may involve cues that rather adversely affect the accuracy of English spelling for Spanish-English bilingual students. For instance, some English phonemes are not represented in Spanish phonology, but there exist nearby phonemes in Spanish that could serve as allophones. This phenomenon indeed explains common English misspellings among Spanish-English bilingual students, such as writing *fonny* for *funny* (Bahr et al., 2015), where the English phoneme /ʌ/ is not present in Spanish, but the letter and sound /o/ can serve as its allophone in Spanish (Wolters & Kim, 2024). Word Dictation items reflecting such phenomenon, if any, could lead Spanish-English bilingual students to score lower on those particular items compared to English-monolingual students with a similar level of overall English transcription skills, thereby indicating DIF items favoring the English-monolingual student group than the Spanish-English bilingual student group.

Current Study

The present study aims to assess the validity and fairness of English writing CBMs when used with linguistically diverse students identified as having intensive early writing needs. By

doing so, this investigation aligns with the objectives of Stage 1 CBM research (Fuchs, 2004). Specifically, I examine whether Word Dictation, a CBM task designed to assess English transcription skills at the word level, exhibits measurement invariance across multilingual and English-monolingual student groups. Recognizing that the psychometric characteristics of assessments are tied to their scores and subsequent interpretations, rather than being inherent properties of the assessment tool itself (AERA/APA/NCME, 2014; Espin & Deno, 2016), I assess measurement invariance using two widely used scoring metrics: WSC and CLS. Furthermore, given that measurement variance at the item level does not necessarily imply measurement variance at the assessment level, I assess both DIF and DTF.

The specific research questions are as follows: (1) To what extent does CBM Word Dictation show measurement invariance *at the item level* concerning the language status of students with intensive writing needs, when analyzing both (a) dichotomous (WSC) and (b) polytomous responses (CLS)? To address this question, I considered multiple analysis methods, including IRT-based and CTT-based approaches, because DIF is a complex phenomenon requiring comprehensive examination, and there is no single technique for its detection. (2) Does measurement invariance hold *at the assessment level*?

Central Hypothesis

Regarding the first research question, I do not hold a strong hypothesis regarding whether and which specific items would function differently for multilingual and English-monolingual student groups with a similar level of latent ability. However, when comparing WSC and CLS scores, I do expect that, among the items identified as displaying DIF with WSC, fewer will be flagged as DIF when using CLS. Some of the DIF detected with WSC may be mitigated under

the CLS scoring method. This expectation is based on both theoretical and methodological considerations. Existing literature suggests that CLS is particularly sensitive to capturing early writing skills among young students struggling with writing and is considered robust (Choi et al., 2023; Hampton & Lembke, 2016; Lembke et al., 2003). Since CLS assigns partial credit for partially correct item responses, it may provide a more accurate representation of a student's overall English transcription skills, making it less prone to DIF. From an analytic standpoint, the method of evaluating measurement invariance with CLS employs an omnibus polytomous DIF statistics that considers the probability of DIF across collapsed score categories within an item. Within this framework, I do not anticipate instances where DIF is identified in CLS, when all score categories are considered, but not in WSC. Instead, I anticipate the opposite pattern—where DIF is found in WSC but mitigated in CLS.

For the second research question focusing on assessment-level analysis, I hypothesize that Word Dictation, in its entirety, would demonstrate measurement invariance across multilingual and English-monolingual student groups. I assume that any DIF effects, if detected, would not systematically favor or disadvantage a specific group. Instead, I expect that such effects, if identified, would balance out at the entire assessment level.

Terminology

No single term can fully encompass the unique language characteristics and identities of students who are learning English as L2. However, in this dissertation, I use the term *multilingual students*, as it recognizes and values what the student possesses rather than emphasizing what they may lack (Helman et al., 2016). I specifically focus on those in the U.S. who are developing or have developed English. These students can be viewed as linguistic

minorities, given that English, their non-dominant language, is the language predominantly used in the society they live in, the schools they attend, and the tests they are given.

I acknowledge that using the term multilingual students without specifying linguistic characteristics, especially their native languages, might reinforce the norm of homogeneity; I proceed with caution and refrain from applying the generalized description in this response to each individual within this diverse population. Additionally, while the current study focuses on whether students use more than one language when referring to multilingual students, it should be noted that linguistic diversity goes beyond the mere number of languages an individual speaks or their specific home languages. Linguistic diversity includes variations in how the same language is used across different contexts, locations, and with varying social consequences (e.g., Adriansen et al., 2022). For students who share the same native language, there exists an extreme within-group variation as a function of L1 exposure, onset of L2 exposure, language proficiency in both L1 and L2, and so on (Thordardottir, 2011).

Chapter 2: Literature Review

Introduction

In this chapter, I present the findings from a systematic review that examines existing studies on the technical adequacy of CBMs in assessing English writing skills among multilingual student populations. The chapter is organized into several sections: an introduction outlining the background and rationale for this literature review, a methods section detailing the literature search process and data coding, a results section presenting the technical evidence found, and a discussion dedicated to interpreting these findings within the scope of the review.

Complexity, Accuracy, and Fluency Framework for Categorizing Writing CBM Scoring Methods

The diverse range of writing CBM scoring metrics can be categorized in different ways. One approach is to consider the specific dimension of writing that each metric is intended to capture. Wagner and colleagues (2019) endorsed this approach by adopting the Complexity, Accuracy, and Fluency (CAF) framework. Researchers have originally introduced the CAF framework to explain L2 acquisition (Housen & Kuiken, 2009) and also applied it to writing (Puranik et al., 2008; Woolpert, 2016). According to the framework and within the context of L2 acquisition, *accuracy* pertains to the skill of producing error-free speech (Lennon, 1990), where errors signify deviations from the norm (Wolfe-Quintero et al., 1998). *Fluency* pertains to the capacity to process the L2 at a pace resembling that of native speakers (Lennon, 1990), or the extent to which the language produced during a task exhibits pauses, hesitations, or revisions (Ellis, 2003). *Complexity*, being the most intricate and ambiguous dimension within the CAF

framework, can be conceptualized as the degree to which the language produced during a task is elaborate and diverse (Ellis, 2003).

Adopting the CAF framework, CBM writing scoring metrics can be classified as follows. First, scoring procedures intended to capture *fluency,* often called production-dependent indicators (Jewell & Malecki, 2005; Keller-Margulis et al., 2016), include WW, WSC, CWS, CLS, and correct punctuation (CP). When using WW, the scorer simply counts the number of words in the student's writing sample. WSC involves counting correctly spelled words in the writing sample, regardless of their appropriateness, similar to how a computer spell-checker assesses correctly spelled words. CWS refers to two adjacent words that are both spelled correctly and used appropriately in a sentence, while CLS refers to two adjacent letters correct in order within a word. CP indicates the number of punctuation marks correctly used at the end of sentences (Keller-Margulis et al., 2016).

Next, scoring indicators designed to assess *accuracy* (i.e., production-independent indicators; Jewell & Malecki, 2005; Keller-Margulis et al., 2016) focus on measuring precision regardless of the quantity of writing generated by the student. These indicators encompass the percent WSC (%WSC) and percent CWS (%CWS). Percent WSC is determined by dividing the number of WSC by the total number of WW. Percent CWS is determined by dividing the number of CWS by the total number of word sequences.

Some scoring procedures are intended to capture *both fluency and accuracy* (i.e., accurate-production indicators; Jewell & Malecki, 2005; Keller-Margulis et al., 2016), such as CIWS and %CIWS. Intended to capture the complexities in writing, CIWS adheres to the criteria for CWS scoring, but it also takes into account the number of incorrect word sequences, which is

then subtracted from the total CWS. Other examples include correct minus incorrect words (CIW), incorrect word sequences (IWS), incorrect letter sequences (ILS), correct minus incorrect letter sequences (CILS), and their variations using percentage scores. Compared to the aforementioned fluency- or accuracy-focused scoring procedures, these methods introduce a heightened level of complexity in the scoring process.

Scoring procedures focusing on the *complexity* of writing are relatively limited. One example is the terminal unit (T-unit), which refers to an independent clause along with any related dependent clauses. T-unit scoring has often been considered challenging for teachers and less promising for writing CBMs than other scoring methods (Campbell et al., 2013; McMaster & Espin, 2007); however, a more recent study (Reno & McMaster, 2024) has validated variations of T-unit that can serve as complementary scoring metrics for sentence-level CBM writing tasks, showing its potential. Additionally, metrics such as type token ratio (CTTR; Carroll, 1964), number of different words (NDW; Scott, 2009), and CWS per item response (CWSR; Wagner et al., 2019) also capture the complexity dimension and have been empirically tested with writing CBMs. Both CTTR and NDW focus on vocabulary. CTTR is a modification of the original measure of vocabulary diversity known as type token ratio (TTR) that calculates the proportion of different words (types) to total words (tokens) in students' writing (Malvern & Richards, 2002; Olinghouse & Leaird, 2009). In NDW, the scorer simply counts the number of different words written. In CWSR, the scorer calculates the average number of CWS per item response, capturing both syntactic complexity and accuracy.

Current Literature Review

The purpose of this literature review was to summarize existing empirical studies that specifically evaluate the reliability and validity evidence of various CBM tasks in assessing English writing skills among multilingual students. In doing so, I paid special attention to different scoring approaches, each capturing different dimensions of writing. I particularly followed the approach of Wagner and colleagues (2019) to categorize CBM writing scoring indicators according to the CAF framework and used this classification to explore potential variations in the reported reliability and validity evidence of writing CBMs for multilingual students. Furthermore, I placed particular emphasis on the diverse L1 backgrounds of the participants across the included studies.

The literature review was guided by the following research questions: (1) To what extent does the literature provide technical adequacy data for writing CBMs in assessing the English writing skills of multilingual students? What specifically does the evidence reveal in terms of reliability, criterion validity, and diagnostic accuracy validity? (2) Does the use of the CAF framework to classify writing CBM scoring methods reveal any variations within this evidence (reliability, criterion validity, and diagnostic accuracy validity)?

Method

Study Eligibility Criteria

To be eligible for this review, each study had to meet the following criteria. First, the participants included multilingual students from kindergarten to Grade 12 learning English as an additional language. Studies that included both English-monolingual and multilingual students were considered when the authors reported writing CBM technical adequacy findings with multilingual students in a disaggregated way. Second, studies administered CBMs with

multilingual students to assess their English writing skills. Third, studies examined at least one of the psychometric attributes of English writing CBMs among reliability (e.g., alternate form, test-retest, internal consistency, inter-scorer), correlational criterion validity, and diagnostic validity (sensitivity, specificity, Area under the Curve [AUC]). I did not impose any constraints on publication type, the location of the studies, and date range. However, the studies had to involve primary data collection and be reported in English.

Literature Search Procedures

I conducted searches in electronic databases of the Education Resources Information Center (ERIC), Academic Search Premier, Education Source, PsycINFO, ProQuest Dissertations & Theses, Psyarxiv, and Edarxiv. I used a range of search terms in the categories of CBM, writing, and multilingual students (see Table 1 for the complete list). Both the search terms and available subject headings (e.g., writing composition, basic writing, curriculum-based assessment, multilingualism) were used to obtain the desired literature. The search terms and strategies were refined through consultations with a university librarian. I performed the electronic database searches on April 13, 2023. I conducted backward and forward citation tracking on April 17, 2023 to identify any additional references. Scopus was used for forward citation tracking.

The initial search, along with the backward and forward citation tracking, resulted in 138 records. After removing the duplicates, I reviewed 82 unique titles and abstracts to identify and remove studies that clearly did not meet the eligibility criteria (e.g., studies that did not use English writing CBMs). I performed deduplication and title/abstract screening using the online screening tool Rayyan QCRI for Systematic Reviews. Next, I reviewed the full texts of 22

records. A total of seven articles met the inclusion criteria and were included in this review. Figure 1 presents the overview of the literature search and selection process.

To establish the reliability of the literature selection process, a subset of 30% (n = 25) of records identified through the database searches and citation tracking were screened by a graduate student in special education. The number of agreements on inclusion decisions was divided by the number of agreements plus disagreements and multiplied by 100. Interrater agreement (IRA) for title/abstract screening was 92%. Disagreements were discussed and resolved. Additionally, 30% (n = 7) of full texts were reviewed by the same graduate student. The IRA for the full-text screening was 100%. Moreover, the coding of the study characteristics (see the following section) was reviewed to ensure the accuracy of the reported information.

Procedures of Study Coding and Quality Appraisal

For each of the seven studies that met the inclusion criteria, I coded the information based on the following categories (see Table 2 for coding descriptors): multilingual students, characteristics of English writing CBMs used (task/prompt type, targeted writing level, administration time, scoring procedure), and criterion measures. In addition, to assess the methodological quality of the included studies, I developed a rubric by integrating components from established evaluation frameworks designed for correlational research (Thompson et al., 2004), as well as for reliability and validity research (Kottner et al., 2011; Lucas et al., 2010; Whiting et al, 2003). Moreover, I integrated elements extracted from a quality appraisal framework used in a pertinent literature review that investigated the validity evidence of reading CBM (oral reading fluency) among multilingual students from kindergarten to eighth grade (Newell et al., 2020). Detailed quality indicators featured in the rubric can be found in Table 3.

To establish IRA on quality appraisal, 30% (n = 3) of the included studies were selected at random. The graduate student, who participated in the IRA process for the literature selection, independently coded the quality indicators for the studies that I had initially coded. IRA was 95%, and all disagreements were discussed and resolved.

Results

Below, I present the results of the systematic review, which was conducted to examine (a) the extent to which existing literature provides technical adequacy data for writing CBMs in assessing English writing skills among multilingual students and the specific findings regarding reliability, criterion validity, and diagnostic accuracy; and (b) whether applying the CAF framework to classify CBM scoring methods highlights any variations in such technical evidence.

Studies Investigating Technical Adequacy of English Writing Curriculum-Based Measurements with Multilingual Students

Availability and Quality of Studies

A total of seven studies were eligible for this review, comprising six peer-reviewed journal articles and one dissertation. Table 4 presents the results of the quality indicator scoring. All seven studies sufficiently discussed the practical relevance of their findings on the reliability and validity evidence of writing CBMs. All studies explicitly reported the reliability and/or validity estimates for all measured variables. Moreover, they all achieved and reported interscorer reliability and/or fidelity of implementation for writing CBMs, surpassing the minimum requirement of 80%. Most studies (n = 5; 71.4%) provided prior evidence supporting the validity evidence and other psychometric properties of writing CBMs.

The description of the criterion measure was insufficient in many studies, as compared to the comprehensive description of writing CBMs. Whereas four studies (57.1%) adequately described the validity evidence and/or other psychometric properties of the criterion measures, two studies (28.6%) did not provide any validity evidence for the criterion measures used. Regarding the description of participants' linguistic background and the instructional context, all seven studies only partially fulfilled the requirement. Specifically, although the studies reported individual participants' language background, particularly their L1s, none of them measured or reported the students' L1 proficiency. Concerning language instruction, while all studies mentioned the model of language instruction provided for the participants, typically using simplified terms (e.g., English as a Second Language [ESL] classroom, bilingual language development program), they often did not provide details regarding the content, duration, and frequency of such instruction. In most cases (n = 6; 85.7%), the administration and scoring process for writing CBMs and criterion measures were described, although the descriptions could be clearer to support replicability. Specifically, important information regarding the demographic characteristics of the administrators, such as their languages, was lacking. While the timing of administering both writing CBMs and criterion measures was reported in the majority of studies (n = 5; 71.4%), there was no explicit mention of whether the time interval between the two administrations was reasonable enough to ensure consistent participation conditions. When it came to discussing and addressing the assumptions of the analysis, such as linearity, independence, normality, or homoscedasticity, none of these assumptions were measured, or at least not reported in four studies (57.1%).

Descriptive Information for Multilingual Students

Table 5 provides a comprehensive description of the characteristics of the studies. The studies covered a range of grade levels, with four studies (57.1%) conducted at the elementary level and three studies (42.9%) at the secondary level. In most of the studies, the student sample involved different L1s, except for one study (Landis, 2019), which focused exclusively on Spanish-speaking students. Across the studies, Spanish was the most frequently found L1; in three studies, Spanish was the L1 for either all of the students or for the largest proportion of the students. In one study (Keller-Margulis et al., 2016), the authors did not explicitly state students' L1s; instead, they reported the students' racial/ethnic identities, with over 90% of the students identifying as Hispanic. I categorized the characteristics of this study under the "Spanish as L1" category. Following that, Somali was the L1 for the largest proportion of students in the two studies. One study (Campbell et al., 2013) did not specify the students' L1s; instead, the authors described that the students' L1s belonged to a group of African languages (Amharic, Eritrean, Oromo, or Somali). Other L1s represented across the studies, although in small proportions, include Arabic, Burmese, Chinese, Czech, Hmong, Karenni, Kirundi, Korean, Laotian, Yoruba Nigerian, Tagalog, Tibetan, Tigrinya, Vietnamese, and Swahili.

Regarding the students' language proficiency, the majority of the studies (n = 5; 71.4%) provided information about their English proficiency to some extent (typically using proficiency labels such as beginning, intermediate, and advanced, as determined by district standards). In contrast, none of the studies provided information on the students' L1 proficiency. Compared to elementary-level studies that reported English proficiency (Smith & Lembke, 2021, 2022), secondary-level studies typically indicated relatively higher proficiency levels, ranging from moderate to high or advanced (Campbell et al., 2013; Espin et al., 2008). All of the studies

described the language of instruction that students received but with varying levels of detail. Three studies (42.9%) reported that students were enrolled in ESL classrooms. Two studies (Keller-Margulis et al., 2016; Landis, 2019) indicated that students were in classrooms designed to support bilingual language development in both English and Spanish, while two other studies (Smith & Lembke, 2021, 2022) indicated English-only classrooms. Information on the instructional time or content was generally limited. Keller-Margulis et al. (2016) provided relatively detailed information, reporting that 75% of classroom instruction was delivered in English.

Characteristics of Writing Curriculum-Based Measurements

In the majority of the studies, English writing CBMs were administered for a brief duration of no more than 5 minutes. Yet, there were two exceptions, with Campbell et al. (2013) allowing for a 7-minute duration and Espin et al. (2008) permitting a 10-minute timeframe; both studies were conducted at the secondary grade level. In the elementary-level studies, a variety of writing tasks or prompts were used, including word dictation (Smith & Lembke, 2022), where students write words dictated by the administrator, measuring word-level transcription skills; picture word prompts (Smith & Lembke, 2021), composed of words accompanied by pictures, targeting sentence-level transcription and text generation skills; and narrative (story) prompts (Keller-Margulis et al., 2016) or descriptive prompts (Landis, 2019), which focus on discourse-level text generation skills. Among the secondary-level studies, one study (Campbell, 2010) used a passage copying task. Other studies used a range of prompts aimed at measuring discourse-level text generation skills, including narrative prompts (Espin et al., 2008) and a combination of narrative, expository, and picture prompts (Campbell et al., 2013).

All the studies employed a variety of scoring procedures. Categorizing these scoring procedures according to the CAF framework, the methods focusing on *fluency* were used in most studies; the most frequently used metrics were WW (often called total words written [TWW]) (*n* = 6, 85.7%), WSC (often termed words written correctly [WWC] or correct words [CW]) (*n* = 5, 71.4%), and CWS (*n* = 5; 7.4%). To capture *accuracy*, studies frequently used %WSC (%WWC or %CW) (*n* = 4, 57.1%) and %CWS (*n* = 3, 42.9%). Less frequently, researchers also used CP and CLS (*n* = 1, 12.5%, respectively) for fluency and %CLS (*n* = 1, 12.5%) for accuracy. In addition, some researchers adopted more complex scoring procedures to capture *both fluency and accuracy.* Among these, CIWS (*n* = 5, 71.4%) was the most commonly used, followed by CIW, IWS, ILS, CILS, WW+CIWS, and %CIWS, each used in one study. Last, only a small number of studies incorporated scoring metrics focusing on *complexity* (the number of T-units and words divided by T-units in Campbell et al., 2013; CWSR in Smith & Lembke, 2021; NDW and CTTR in Landis, 2019).

Criterion Measures

A variety of English criterion measures were used to examine the correlational criterion validity evidence of English writing CBMs. Studies included state-developed assessments (e.g., Minnesota Basic Skills Test [MBST] in Campbell, 2010 and Campbell et al., 2013; STAAR in Keller-Margulis et al., 2016), commercially developed assessments (e.g., Test of Written Language-Third Edition [TOWL-3] in Campbell et al., 2013), English language proficiency assessments (e.g., ACCESS in Smith & Lembke, 2021, 2022; Test of Emerging Academic English [TEAE] in Campbell, 2010 and Campbell et al., 2013; English Language Proficiency Assessment for the 21st Century [ELPA21] in Landis, 2019), and teachers' holistic ratings of

writing quality (Campbell, 2010; Campbell et al., 2013) as criterion measures. Often times, these assessments included subtests in various academic domains, and studies explored correlations between writing CBMs and different subtest scores, not limited to writing. While presenting the correlational validity evidence, I specifically focused on correlations examined in relation to writing scores. Studies involving various subtest scores are recorded as table notes in Table 5.

Findings on Reliability and Validity

Table 5 includes reliability (alternate-form reliability, inter-scorer reliability) and validity (criterion validity, diagnostic accuracy validity) data reported within the studies. I presented the range of coefficients for each scoring method whenever possible, rather than providing a single aggregated coefficient. The range reflects the variations in writing CBM administration times (e.g., spring, fall, winter), administration durations (e.g., 3, 5, 7, 10 minutes), types of writing prompts (e.g., narrative, expository, pictures), and the use of a variety of criterion measures (e.g., teachers' ratings, commercially- or state-developed standardized tests), particularly in the case of criterion validity. Regarding alternate-form reliability coefficients (measured as Pearson's *r* correlation), the values can be interpreted as follows: $r > .80$ as indicating a relatively strong correlation, .70 to .80 as moderately strong, .60 to .70 as moderate, and < .60 as weak (McMaster & Espin, 2007). Concerning inter-scorer reliability, I considered a percent agreement of at least 80% acceptable, based on at least 20 percent of the judgments, following the criteria set by What Works Clearinghouse (WWC, 2022). In terms of diagnostic accuracy validity, despite variations in recommendations regarding acceptable levels of sensitivity and specificity, I interpreted sensitivity and specificity values exceeding .70 as acceptable, aligning with the standards used in studies evaluating reading CBMs for screening (Kilgus et al., 2013; Silberglitt & Hintze, 2005).

Alternate-Form Reliability. Six studies (85.7%) examined alternate-form reliability. These studies collectively reported 37 distinct correlation coefficient ranges using various scoring methods. Across these studies, the coefficients spanned from a minimum value of $r = .02$ (36 students primarily with African languages as their L1s, Campbell et al., 2013) to a maximum of .98 (73 students primarily with Spanish as their L1, Smith & Lembke, 2022), indicating weak to relatively strong alternate-form reliability. The substantial variability in the overall range of coefficients particularly stemmed from one study (Campbell et al., 2013), in which the authors reported minimum values of .02 (for %CW) and .13 (for Words/#T-units) within the ranges.

Inter-Scorer Reliability. Out of the six aforementioned studies, three (Landis, 2019; Smith & Lembke, 2021, 2022) also included inter-scorer reliability coefficients. Smith and Lembke examined inter-scorer reliability on a randomly chosen 24% subset of the writing samples from the students using Spanish as their L1 and calculated the ratio of total scoring agreements to the sum of both agreements and disagreements. The outcomes ranged between 88% and 98% in Smith and Lembke (2021), and 94% to 99% in Smith and Lembke (2022), meeting the WWC standards. Landis (2019), on the other hand, reported an intraclass correlation coefficient (ICC) value of .96 for a randomly selected 20% of the writing samples from students with Spanish as their L1.

Criterion Validity. All of the studies included in this review presented criterion validity data (measured as Pearson's *r*). Across these studies, the coefficients ranged from a negative value (-.13) to a maximum of .84. Both of these extremes were obtained from Campbell et al. (2013), in which 36 students in Grades 10-12 with unspecified African languages as their L1s participated. Furthermore, out of the 47 distinct ranges of correlations reported using different

scoring methods across the studies, 27 of these ranges either exceeded the minimum value above .50 or included it within their range.

Diagnostic Accuracy Validity. Two studies (28.6%) provided data on diagnostic accuracy validity. First, Keller-Margulis and colleagues (2016) reported sensitivity (the probability that a student scoring below the writing CBM cut score actually did not pass the state test) and specificity (the probability that a student scoring above the writing CBM cut score successfully passed the state test), as well as AUC values, representing the overall classification accuracy. Focusing on students with Spanish as their L1, they reported diagnostic validity data for English Language Learners (ELLs) and Monitored students who had exited the services. In terms of selecting cut scores, these researchers identified and used writing CBM scores that optimized sensitivity while retaining specificity levels at .70 or higher. Their findings indicated that sensitivity and specificity values were 1.0 and 0.73 for ELL students, respectively, when using %CWS; these values represent acceptable levels. For Monitored students, the achieved sensitivity was 1.0 with %CWS, %WSC, and CIWS, with the corresponding values of specificity of .90, .86, and .93, also representing acceptable levels. However, certain scoring metrics (CIWS for ELLs, CP for Monitored students) failed to establish a cut point that offered at least .70 sensitivity and specificity. Regarding AUC values, Keller-Margulis and colleagues reported a range of .76 to .84 for ELLs and .86 to .93 for Monitored students. However, none of these AUC values were statistically significant for ELLs and Monitored student groups.

Next, Landis (2019), who focused on Spanish-English bilingual students, presented AUC values. Landis used a range of cut scores, including those for overall English proficiency classification and domain-specific 20th percentile scores in writing, reading, speaking, and

listening on the ELPA21; I focused on the values derived from the 20th percentile writing score. The author reported AUC values spanning from .60 to .81 for NDW and .55 to .80 for CTTR. However, the overall pattern of ROC outcomes did not consistently indicate adequate classification accuracy for NDW and CTTR.

Reliability and Validity Data Categorized Using the Complexity, Accuracy, and Fluency Framework

To address the second research question, I categorized the writing CBM scoring methods according to the CAF framework as fluency-focused, accuracy-focused, fluency-and-accuracy-focused, and complexity-focused scoring procedures. I then presented the reliability and validity data obtained from these scoring methods for each category (see Table 6). Note that, regarding reliability, I specifically focused on alternate-form reliability evidence, as the studies did not provide disaggregated inter-scorer agreement data for distinct scoring methods.

Fluency-Focused Scoring Methods

The alternate-form reliability of the measures displayed a range of strengths: WW (*r*s = .55–.97; across 5 studies); WSC (*r*s = .58–.97; across 5 studies); CWS (*r*s = .60–.94; across 4 studies); ILS (*r*s = .61–.97; in 1 study); and CLS (*r*s = .91–.97; in 1 study). Regarding criterion validity, the correlations spanned as follows: WW (*r*s = -.05–.66; across 7 studies), WSC (*r*s = -.00–.81; across 7 studies), CWS (*r*s = .08–.82; across 6 studies), CP (*r*s = -.04–.50; across 2 studies), and CLS (*r*s = .56–.78; in 1 study). Diagnostic validity data for fluency-focused scoring methods were not available.

Accuracy-Focused Scoring Methods

Alternate-form reliability when assessed using accuracy-focused scoring methods showed a wide range as well: %CWS (*r*s = .42–.88; across 3 studies); %WSC (*r*s = .02–.95; across 4 studies); and %CLS (*r*s = .78–.93; in 1 study). Criterion validity for these methods produced the following results: %CWS (*r*s = -.02–.82; across 4 studies), %WSC (*r*s = -.06–.82; across 5 studies), and %CLS (*r*s = .47–.60; in 1 study). Diagnostic validity data were available for %CWS (sensitivity at 1.0, specificity ranging from .73 to .90, ACU values of .84 and .90) and %WSC (sensitivity at 1.0, specificity at .86, AUC of .86).

Fluency-and-Accuracy-Focused Scoring Methods

For specific scoring methods that capture both writing fluency and accuracy together, the alternate-form reliability coefficients exhibited the following ranges: WW+CIWS (*r*s = .59–.90; in 1 study); CIWS (*r*s = .30–.91; across 4 studies); CIW (*r*s = .65–.69; in 1 study); and CILS (*r*s = .86–.97; in 1 study). Regarding criterion validity, the correlation coefficients ranged as follows, using a range of criterion measures: CIW (*r*s = .33–.64; in 1 study), CIWS (*r*s = .00–.84; across 6 studies), CILS (*r*s = .51–.75; in 1 study), and WW+CIWS (*r*s = .41–.83; in 1 study). Diagnostic accuracy validity data were available only for CIWS, with sensitivity at 1.0, specificity at .93, and AUC of .93.

Complexity-Focused Scoring Methods

Reliability coefficients for complexity-focused scoring methods ranged as follows: W/T-units (*r*s = .13–.62; in 1 study); CTTR (*r*s = .51–.65; in 1 study); CWSR (*r*s = .30–.94; in 1 study); #T-units (*r*s = .53–.82; in 1 study); and NDW (*r*s = .62–.72; in 1 study). As for criterion validity, the results were as follows: NDW (*r*s = .24–.48 in 1 study), CTTR (*r*s = .20–.49; in 1 study), CWSR (*r*s = -.10–.56; in 1 study), #T-units (*r*s = .05–.58; in 1 study), and W/T-units (*r*s

= -.13–.52; in 1 study). Diagnostic accuracy validity data were available for NDW (AUC .60–.81) and CTTR (AUC .55–.80), but the diagnostic performance was not meaningful.

Discussion

In this systematic review, I aimed to comprehensively examine the extent to which the existing literature provides technical adequacy data for writing CBMs in assessing English writing skills among multilingual students and to specifically analyze what these data reveal regarding different types of reliability and validity. In addition, by adopting the CAF framework and categorizing the various writing CBM scoring methods based on the CAF dimensions, I sought to identify potential variations, if any, in the reported outcomes of reliability and validity by the studies. Below, I discuss overall findings related to the research questions and consider potential implications for future research and practice.

Technical Adequacy of Writing Curriculum-Based Measurements in Assessing Multilingual Students' English Writing Skills

Availability and Quality of Studies

This review revealed a surprisingly small number of studies (N = 7) that have investigated and provided technical adequacy data for CBMs when assessing English writing skills in multilingual student populations (K–Grade 12). This notably limited volume of research stands out as inadequate when contrasted with the over 20 studies found in relatively recent meta-analyses (Romig et al., 2017, 2021) and the systematic review conducted over a decade ago (McMaster & Espin, 2007) that involve broad student populations.

Additionally, the findings of this review suggest the technical adequacy evidence for many writing CBMs is established for students whose L1s include Spanish, Somali, and African

languages. Yet, these languages only represent a fraction of the diverse range of L1s found in multilingual student populations in U.S. public schools (National Center for Education Statistics, 2023). Therefore, continued research is necessary to assess the technical adequacy of English writing CBMs for student populations using L1s not covered in the reviewed studies.

Overall, the identified studies exhibited satisfactory quality in terms of methodology and the sufficiency of reported information. However, one crucial aspect, which is especially vital in research centered on multilingual student populations, was not adequately addressed in many of the included studies. This aspect pertains to the students' language characteristics and the language of instruction provided. The quality appraisal in the current review revealed that, although the studies did report the multilingual students' L1s, they often omitted essential details, such as the students' proficiency in their L1 and the extent or duration of their exposure to the English language. Moreover, in terms of language instruction, the studies often lacked this information or relied on simplistic descriptions (e.g., "ESL class" or "bilingual program") without providing in-depth insights, such as the proportion of L1/L2 usage and the frequency and duration of language instructional activities, if applicable. Considering that students' language proficiency and their knowledge of writing-related concepts in both their L1 and additional language(s), language use at home, and educational backgrounds, among other factors, can influence their development in L2 writing and should be considered within the assessment context (Genesee et al., 2006), it would be essential to include such information in research on the use of writing CBMs with multilingual student populations.

Technical Adequacy Evidence Presented in Studies

Alternate-Form Reliability. Six out of seven studies investigated the reliability of writing CBMs using alternate forms, together yielding correlation coefficients that ranged from *r*s = .02 to .98 across various scoring procedures. This wide range of correlations indicates different levels of alternate-form reliability, spanning from weak to relatively strong; however, a more detailed analysis of the findings reveals that out of the 37 ranges, 31 of them either included the value of .70 (the minimum requirement for assessments used for progress monitoring purposes; Salvia et al., 2017) within their range or exceeded it. Furthermore, 25 of these ranges either surpassed the minimum value of .80 (the minimum requirement for assessments used for screening decision purposes; Salvia et al., 2017) or incorporated it within their range. Although direct comparisons are not possible, this finding generally aligns with prior research that has reported alternate-form reliability estimates for standardized writing measures ranging from .70 to above .90 (Taylor, 2003). This finding is also in line with findings from the previous synthesis (McMaster & Espin, 2007) which observed alternate-form reliability coefficient ranges of .006 to .96 among students at elementary and secondary grade levels. Taken together, with caution, this review's findings suggest that certain scoring metrics exhibit stronger evidence of alternate-form reliability for writing CBMs than others.

Inter-Scorer Reliability. Inter-scorer reliability, as compared to alternate-form reliability, has been less frequently examined, with only three studies in this review addressing it. The reported percentage agreement ranged from 89% to 99%, suggesting a satisfactory level (WWC, 2022). However, besides inter-scorer reliability and alternate-form reliability, this review identified a lack of evidence concerning other types of reliability, such as internal consistency reliability and test-retest reliability. Researchers have indicated that measurement

errors in writing CBMs can originate from various sources (Winkes & Schaller, 2022) and that these measurement errors may be even more pronounced for multilingual students due to extraneous variables such as cultural and linguistic backgrounds among others (Sandberg & Reschly, 2011). Incorporating various forms of reliability statistics is beneficial as each of them can address different aspects; for example, internal consistency and test-retest reliability can help detect measurement errors stemming from poorly worded questions, or in the case of writing CBM, instructions or prompts (WWC, 2022). Thus, expanding the scope of research on the reliability of writing CBM for multilingual student populations would be essential.

Criterion Validity. All seven studies included in this review investigated the correlations between the performance of multilingual students in writing CBMs and various criterion measures in writing. Together, the studies yielded a broad spectrum of correlations, ranging from negative (-.10) to relatively strong (.81); I discuss potential reasons for this broad range in the next section. The previous synthesis (McMaster & Espin, 2007) similarly identified a broad range of criterion validity coefficients (*r*s = -.24–.99) for writing CBMs within a diverse student population. Romig and colleagues (2017), in their meta-analysis, found an average criterion validity coefficient of *r* = .55 for writing CBMs across various scoring methods. Although the findings regarding criterion validity in this review generally align with previous literature, as more studies accumulate in this research area, future researchers might consider synthesizing the correlational validity outcomes through a meta-analysis and then comparing them with findings from the earlier meta-analysis.

Diagnostic Accuracy Validity. Diagnostic accuracy data is especially vital for screening tools as it informs determining the cut scores (thresholds) for classifying students needing

additional support (Jenkins et al., 2007; Kilgus et al., 2014). This review uncovered a significant lack of such data for multilingual students. Only two studies examined the diagnostic accuracy of English writing CBMs in multilingual students, with one study (Landis, 2019) observing limited meaningful diagnostic accuracy for NDW and CTTR. More research is needed to evaluate the diagnostic accuracy of English writing CBMs with linguistically diverse students, which is especially crucial given the growing acknowledgment of the importance of incorporating writing measures into screening procedures from early on (e.g., Kim & Petscher, 2023).

Variability in Technical Adequacy Based on Complexity, Accuracy, and Fluency Framework Categorization

The application of the CAF framework did not reveal noticeable variations in technical adequacy data *across* the dimensions (fluency-focused, accuracy-focused, fluency-and-accuracy-focused, complexity-focused scoring procedures), at least partly due to the limited number of studies included in the review. Nonetheless, by applying the framework, the current review did reveal some variations in the technical adequacy data *within* each dimension, particularly in terms of criterion validity data, and I discuss them in detail below. Although it is important to approach the interpretations with caution and validate them through continued research, these findings could provide practical insights, such as assisting educators in making informed decisions about which specific scoring methods to consider within a particular dimension.

Fluency-Focused Scoring Methods

Among the scoring methods that focus on writing *fluency*, CLS demonstrated relatively stronger technical adequacy evidence for multilingual student populations, particularly in terms of criterion validity (*r*s = .56–.78), when compared to other methods like WW (*r*s = -.05–.66),

WSC (*r*s = -.00–.81), CWS (*r*s = .08–.82), and CP (*r*s= -.04–.50); however, note that this finding for CLS is based on only one study (Smith & Lembke, 2022). The observed criterion validity ranges for CLS broadly align with those reported in previous research that examined students at elementary grade levels. Earlier studies indicated that criterion validity coefficients for word dictation scored with CLS ranging from .52 to .92 (Lembke et al., 2003) and .48 to .50 (Hampton & Lembke, 2016) in relation to students' written responses to a picture story starter and the Test of Early Written Language-2, respectively. Moreover, literature indicates that CLS can prove especially useful when dealing with words slightly more challenging than a student's current writing level (Frisby, 2016). As such, educators may prioritize this scoring method when assessing the writing fluency of multilingual students, especially those in elementary grade levels or those in the process of building foundational English writing skills.

In contrast to CLS, WW, WSC, and CP included negative values as their minimum criterion validity coefficients, and it is important to understand the contexts in which these less favorable criterion validity outcomes emerged. In this review, I identified the negative criterion validity value for WW in Keller-Margulis et al. (2016); this negative correlation was obtained when the authors examined the relation between the story starter task administered in the spring to fourth-grade students, primarily Spanish-speaking students in their L1, and STAAR writing. Similarly, the negative values for WSC and CP emerged from the same study, but they were observed when the story starter task was administered in the winter and spring, respectively. To gain a deeper understanding of these findings, continued research is necessary to explore whether specific grade levels, language characteristics, or other factors might be linked to the relatively weaker technical adequacy of these scoring methods for multilingual students.

Accuracy-Focused Scoring Methods

When considering accuracy-focused methods, %CLS demonstrated higher criterion validity coefficients (*r*s = .47–.60; again, note that these estimates are based on a single study) compared to %CWS (*r*s = -.02–.80) and %WSC (*r*s = -.06–.82). Negative criterion validity correlations for %CWS and %WSC were once again identified in Keller-Margulis et al. (2016). Smith and Lembke (2021) also observed a negative correlation for %WSC when administering the picture word task to first graders who primarily spoke Spanish as their L1 and correlating it with their ACCESS writing scores.

Fluency-and-Accuracy-Focused Scoring Methods

Among the fluency-and-accuracy-focused scoring methods used in the studies, CIWS was most frequently examined. This scoring procedure, like many others, yielded a wide range of criterion validity coefficients (*r*s = .00 to .84). Despite the generally acceptable criterion validity, negative criterion validity values were identified in Keller-Margulis et al. (2016) and Smith and Lembke (2021) among 4th-grade primarily Spanish-speaking students and 1st-grade primarily Spanish-speaking students, respectively. The presence of these negative values may be linked to specific grade characteristics or the writing levels of the student samples in these studies. CIWS was initially developed primarily for use with secondary students (Espin et al., 2000), and stronger validity evidence for CIWS has been identified among older students within the K-12 grade range (McMaster & Espin, 2007). However, at the same time, Romig and colleagues' meta-analysis (2017), which encompassed various grade levels, highlighted CIWS with the highest mean criterion validity coefficient (*r* = .60) among different scoring metrics, leading the researchers to conclude that the criterion validity of writing CBMs improves with the

increased complexity of the scoring procedure. Further research is warranted to investigate whether this conclusion holds true when assessing English writing skills in multilingual student populations.

Complexity-Focused Scoring Methods

Complexity-focused scoring methods were less frequently explored compared to fluency-, accuracy-, and fluency-and-accuracy-focused scoring methods, resulting in a limited amount of technical adequacy data. While Landis (2019) made strides by investigating vocabulary indices (NDW, CTTR), the provided alternate-form reliability and criterion validity evidence were not as robust. In Smith and Lembke (2021), where the promising scoring method CWSR was used, the criterion validity ranged from -.10 to .56 in relation to ACCESS, with the negative value being observed for first graders. Ongoing research regarding the use of CWSR for multilingual students is of particular importance, as despite its relatively recent development, CWSR exhibited robust validity and reliability in Wagner and colleagues' study (2019) conducted in the context of picture word tasks administered to a broad student population across Grades 1-3.

Limitations

When interpreting this review's findings, it is important to acknowledge the following limitations. First, although I aimed to perform an exhaustive and comprehensive literature search, this review ended up incorporating only a limited number of studies. There is a chance that pertinent studies could have been unintentionally missed. Thus, it is important to approach the interpretation of the findings with caution.

Second, I only focused on studies administering writing CBMs in English, as my emphasis was on assessing the technical aspects of measures that evaluate English writing skills among multilingual students in the U.S. However, this emphasis may divert attention away from studies that evaluated the technical rigor of writing CBMs in students' various L1s. During the literature search process for this review (constrained to studies published in English), I came across only one study (Sanchez, 1995) that explored the technical adequacy of writing CBMs in multilingual students' L1 (Spanish). While I had to exclude this study due to the review's specific focus, it is crucial to acknowledge the significance of adopting a bilingual/multilingual perspective (e.g., Hopewell & Butvilofsky, 2016; Butvilofsky et al., 2021; Flores & Rosa, 2015). Moving forward, further research is needed to investigate writing CBM developed for use in both students' native language and English.

Third, when investigating variations in the technical features of writing CBMs, this review primarily focused on one particular characteristic of the measures, which is a range of scoring procedures. This review did not closely examine potential variations that might arise due to different types of writing prompts, variations in administration time durations, or the various criterion measures used to assess correlational validity. While the previous meta-analysis (Romig et al., 2021) found no distinct pattern in the strength of criterion validity evidence for writing CBMs based on different types of writing prompts (e.g., picture, narrative, expository, text copying, picture-word, picture-story) and varied durations (ranging from 1.5 to 10 minutes), these results might not directly apply to multilingual learner populations. For instance, it is possible that certain writing prompts may contain cultural elements that are more or less familiar to some multilingual students, impacting the technical rigor of the measures. Furthermore,

concerning criterion measures, employing less direct writing measures (e.g., holistic rating) as criterion measures might be associated with lower correlations compared to more direct criterion measures of writing (e.g., writing subtests of standardized assessments) (McMaster & Espin, 2007).

Future Research Directions

Several clear directions emerge from this review. First, although this study prudently demonstrates the satisfactory technical rigor of the existing CBM writing tools when applied to multilingual students, it exclusively addresses certain aspects of validity and reliability evidence. This review does not provide insights into other various aspects of validity evidence, such as structural validity (Messick, 1995). Structural validity examines whether the assessment's internal structure reflects the construct(s) it aims to measure and if that structure remains consistent across different subgroups within the intended population (AERA/APA/NCME, 2014; Messick, 1995). Gathering evidence of structural validity is essential for interpreting a student's scores accurately, which are believed to represent the underlying construct(s).

Second and in relation to the previous point, I was not able to find any prior investigation explicitly focusing on fairness evidence, such as assessing measurement invariance of the measures (e.g., differential item functioning) across multilingual and English-monolingual students. The evidence that the measures demonstrate satisfactory correlations with criterion measures within multilingual students does not necessarily guarantee that the measures function equivalently, without having potential bias against linguistically and culturally marginalized students. Therefore, a more rigorous evaluation of English writing CBMs, with a clear focus on

fairness issues, is required to ensure that these measures offer valid and informative data for linguistically diverse students.

Third, it is important to clarify that the conclusions drawn from this review exclusively validate the technical adequacy of *static* scores obtained from the administration of writing CBMs *at a particular time point.* Out of the three stages of research essential for CBM (Fuchs, 2004), this review only addressed the Stage 1 questions. Although I did attempt to examine studies that investigated the technical attributes of the CBM slopes derived from repeated measurements, none were identified. Future research should tackle Stage 2 questions by examining the reliability, validity, and sensitivity to growth of English writing CBMs to provide evidence supporting their use in progress monitoring for multilingual students.

Conclusion

In this current systematic review, I comprehensively examined available studies focusing on the technical adequacy of CBMs in assessing English writing skills among multilingual student populations. Based on the seven included studies, I presented findings on alternate-form reliability, inter-scorer reliability, criterion validity, and diagnostic accuracy validity. I also applied the CAF (complexity, accuracy, fluency) framework to reveal potential variations in technical adequacy data depending on the writing dimensions that different scoring methods target to measure. Results indicated that English writing CBMs, evaluated through diverse scoring procedures, generally demonstrated sufficient reliability and validity evidence when assessing the English writing skills of multilingual students. While the application of the CAF framework did not reveal noticeable differences in technical adequacy favoring scoring methods targeting a particular dimension, it did reveal variations within each dimension. This review

highlights the need for an expanded body of research in this area, particularly involving students with a diverse range of L1s and involving the evaluation of measurement invariance of the measures for linguistically diverse students. Further work is also warranted to assess the technical adequacy of both static scores and slopes yielded from English writing CBMs for multilingual students.

Chapter 3: Method

Data Source

In the current investigation, I used data drawn from a large, randomized control trial (RCT) (McMaster et al., 2024) that spanned three academic years (2018-19, 2019-20, 2021-22; data were not collected in 2020-21 due to the COVID-19 pandemic). The primary purpose of the RCT was to investigate the effects of a comprehensive professional development program called Data-Based Instruction: Tools, Learning, and Collaborative Support (DBI-TLC; Lembke et al., 2018; McMaster et al., 2020) that was designed to support teachers' implementation of DBI for students with intensive early writing needs. The RCT involved 23 public schools located in two Midwestern states, and three cohorts of teachers and students. In total, 154 elementary teachers, primarily special education teachers, were recruited and randomly assigned to either the treatment or business-as-usual control group. These teachers nominated students who required intensive early writing intervention, specifically those who (a) were in Grades 1–3, though later grades could be considered if they had early writing needs; (b) possessed at least basic English skills and could write one or more letters in English; and (c) received instruction in the general education curriculum. This process led to a total of 523 nominated students who obtained parental consent.

The research team screened these nominated students using two types of CBM writing tasks, namely Word Dictation and Picture Word, to identify target students who were considered to have severe difficulties in early writing. For each teacher, nominated students with the lowest scores in both tasks were selected for inclusion in the RCT. The screening resulted in a total of 377 target students, who received either DBI in early writing (n = 207) from treatment teachers

or business-as-usual writing instruction (n = 170) from control teachers. After accounting for attrition, complete screening, pre- and post-test data were available for 309 students. For further details about the RCT, please refer to McMaster et al. (in press).

Analytic Sample

The analytic sample used for the current investigation comprises students who participated in the initial screening phase and were subsequently qualified as eligible participants in the RCT due to their most intensive writing needs. Additionally, among these target students (n = 377), I specifically focused on those for whom information about their home language was available from a demographic survey completed by school district personnel or classroom teachers during the RCT. The survey included a question about home languages with five response options: English, Spanish, Somali, Hmong, and Other (with an option to specify the "other" language).

Excluding 28 students for whom school personnel did not provide a response to this home language question, the total number of students considered for the current analysis was 349. I categorized students as multilingual if they were reported as using any language other than English (Spanish, Somali, Hmong, or Other) and as English-monolingual if they were reported as using English. For the current study, 67 were identified as multilingual, and 282 were categorized as English-monolingual. Fifty-five (82%) of the 67 multilingual students, and four (1.4%) of the 282 English-monolingual students were reported by their schools to be receiving English Learner services.

Demographic Characteristics

Table 7 presents demographic information for the analytic sample. Most students were male (73.1% of multilingual students, 69.1% English-monolingual students) and in Grades 1–3 (80.6% of multilingual students, 79% of English-monolingual students). A significant proportion of students were receiving special education services (64.2% multilingual and 93.3% English-monolingual students). The sample represented various racial/ethnic categories: for multilingual students, 50.7% Hispanic/Latino(a) American, 23.9% Black/African American, and 14.9% Asian American/Pacific Islander; for English-monolingual students, 70.2% White/European American, 16.3% Black/African American, and 5.3% multiracial. Among the multilingual students, diverse home languages were reported: 33 students (49.3%) indicated Spanish as their home language, 14 (20.9%) reported Somali, one student (1.5%) reported Hmong, and 19 students (28.4%) reported using languages categorized as "other," including Korean, Vietnamese, Russian, Chinese, Amharic, Telugu, Portuguese, and Malayalam.

Chi-square test results (see Table 7) indicated that multilingual and English-monolingual student groups were comparable in terms of some demographic variables but not all. There was no statistically significant difference in terms of sex distribution between the groups. However, significant differences ($p < .05$) emerged in other demographic variables (grade, race/ethnicity, special education eligibility, primary disability category, English Learner service eligibility, and home languages). Variations in English Learner service eligibility, home languages, and race/ethnicity between the groups were anticipated. Regarding grade levels, although specific distributions varied, the majority of students in both groups were in Grades 1–3.

In terms of special education status, a higher proportion of English-monolingual students (93.3%) received special education services compared to multilingual students (64.2%). The

primary disability categories frequently reported for English-monolingual students included other health disability (25.1%), specific learning disability (24.3%), and autism (21.3%), whereas, for multilingual students, autism (32.6%) was the most reported primary disability category, followed by specific learning disability (18.6%) and other health disability (14.0%).

Performance on English Writing Measures

Despite the variations in some demographic characteristics between the groups, most importantly, the two groups were considered comparable in terms of their writing skill levels. Table 8 provides a summary of descriptive statistics of English writing scores that multilingual and English-monolingual students obtained across a range of measures: CBM Picture Word, CBM Story Prompt, and Kaufman Test of Educational Achievement-Third Edition (KTEA-3; Kaufman & Kaufman, 2014). (Note that descriptive statistics of CBM Word Dictation scores, the focus of this study, are presented in Table 10 separately and discussed later.) These measures were all administered before the intervention started. For details about how these measures were used in the RCT, see McMaster et al. (2024). An important note here is that since these measures have not undergone measurement invariance evaluations, DIF items might exist, potentially affecting the score interpretation. Thus, the results and interpretations of the group comparisons below should be considered preliminary.

The average score of Picture Word across two alternate forms (Form A and Form B), when scored by counting correct word sequences, was 7.82 for multilingual students and 7.99 for English monolingual students. These scores fall approximately at the 25th percentile for Grade 1 and below the 10th percentile for Grades 2 and 3, according to established benchmarks for the measure (McMaster & Lembke, 2018). Regarding the Story Prompt, the average scores were

5.08 for multilingual students and 4.11 for English monolingual students, falling between the 25th and 50th percentiles for Grade 1, around the 10th percentile for Grade 2, and below the 10th percentile for Grade 3 (McMaster & Lembke, 2018). In the KTEA-3 Written Language composite scores, the average score was slightly higher for multilingual students (71.85) than English-monolingual students (68.24).

The results of *t*-tests (see Table 8) revealed there were no statistically significant differences in the performance of the multilingual and English-monolingual student groups in any of these measures. Additionally, absolute values of effect sizes, calculated as Hedges' *g*, ranged from .02 to .08 across writing measures, confirming the between-group equivalence. What Works Clearinghouse (WWC, 2022) suggests that satisfying the baseline equivalence standard requires baseline group differences, or effect sizes, no greater than .25 *SD*s on key measures. For differences between .05 and .25 *SD*s, statistical adjustments are required in impact analyses. Differences less than or equal to .05 *SD*s satisfy the baseline equivalence standard and do not require statistical adjustment.

Measure

Word Dictation is designed to assess word-level transcription skills in English, which encompasses spelling and handwriting. Word Dictation is particularly suitable for young students or those who are in the beginning stages of learning to write words (McMaster & Lembke, 2018), as compared to passage-level writing CBM tasks (Campbell, 2010; Ritchey et al., 2016). In the RCT (McMaster et al., 2024), most teachers selected Word Dictation as a progress monitoring tool for their students, further suggesting its suitability for young, beginning writers.

Word Dictation adheres to the CBM assessment framework, merging standardized measurement and traditional psychometrics with elements from behavioral and observational assessment methods including fixed-time recording, repeated performance sampling, and graphical representation of time-series data (Fuchs & Deno, 1991). Word Dictation is administered on a one-to-one basis, where the teacher dictates each word with one repetition, and the student writes down the word. The measure is a time-limited task, with a duration of 3 minutes, meaning that the process of the teacher dictating the word and the student writing it down is allowed to occur as much as possible within the limited time frame.

Word Dictation comprises 20 parallel forms, enabling educators to use it for progress monitoring purposes. Each form comprises 30 items, with each item representing a word. The word lists were created to align with the spelling patterns specified in the Common Core State Standards (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010).

During the screening phase in the large RCT, students completed two forms, Form A and Form B. See Table 9 for the list of words included in Forms A and B. Also, see Appendix A for detailed information on administering Word Dictation. In the RCT, Form A was always administered before Form B; thus, there is a possibility of a test-order effect, such as students feeling more fatigued during Form B or experiencing a greater practice effect in Form B compared to Form A (despite each form starting with a practice item). Students' responses were scored using various metrics, including words written, WSC, CLS, and incorrect letter sequences. For more information on these scoring methods, see Appendix B.

Previous studies demonstrated the technical adequacy of Word Dictation for young students, including those who are multilingual and with intensive writing needs. For students in Grades 1-3, researchers (Hampton & Lembke, 2016; Lembke et al., 2003; Poch et al., 2019) have reported sufficient levels of alternate form reliability coefficients (*r*s ≥ .89 to .95) and criterion validity (*r*s = .11 – .77); in CBM, reliability coefficients greater than .80 are often considered to indicate a strong correlation, .70 to .80 a moderately strong correlation, .60 to .70 a moderate correlation, and coefficients below.60 a weak correlation (McMaster & Espin, 2007). Researchers also identified that Word Dictation is sensitive to capture writing growth over time for young students with intensive writing needs (Choi et al., 2023). Some researchers specifically focused on young students eligible for ELL services and examined evidence of reliability (*r*s = .88 – .97; Smith & Lembke, 2022), as well as concurrent (*r*s = .56 – .81 in Smith & Lembke, 2022; *r*s = .32 – .70 in Keller-Margulis et al., 2016) and predictive validity (*r*s = .61 – .73 in Smith & Lembke, 2022) in relation to standardized norm-referenced measures of written expression in English.

In the context of the present study sample, alternate form reliability and internal consistency reliability of the measure were satisfactory. Among multilingual students (*n* = 67), the correlation (Pearson's *r*) between Form A and Form B ranged from .94 (when using scores based on WSC) to .97 (CLS scores). For English-monolingual students (*n* = 282), the alternate form reliability coefficients ranged from .89 (WSC) to .95 (CLS). Additionally, internal consistency reliability for WSC scores, measured by Cronbach's alpha (Cronbach, 1951), was .99 for Form A and .98 for Form B for the full sample. For the purpose of this study, I used WSC and CLS scores, for reasons described in the "Data Analytic Strategy" section.

Procedures

During the screening phase of the RCT, trained graduate research assistants (GRAs) administered Word Dictation Form A and Form B. The testing was done on a one-on-one basis in a quiet setting at each student's school over a two- to three-week period. For Word Dictation, GRAs demonstrated 99% fidelity to the Word Dictation administration protocol during a mock administration and 98% during actual administrations. Graduate research assistants also scored students' Word Dictation responses. Each GRA reached an interscorer agreement of at least 85% with a project coordinator on two student samples of Word Dictation prior to the testing. The average interscorer agreement for Word Dictation was 99% during the actual administrations. See McMaster et al. (in press) for detailed procedures of administering and scoring Word Dictation.

Data Analytic Strategy

I used R 4.1.2 (R Core Team, 2020) for all analyses. Each form of Word Dictation contains a total of 30 items. For the current analysis, I included a subsample of items: the first 15 items (Items 1-15) for Form A and the first 14 items (Items 1–14) for Form B. This decision was made because students, on average, answered the first 13 to 15 items within 3 minutes, without reaching the later items, resulting in a considerable number of missing responses in later items. Specifically, the mean number of items the students reached, whether correct or not, was 13.82 (*SD* = 6.27) in Form A and 14.42 (*SD* = 6.43) in Form B for multilingual students, and 15.32 (*SD* = 6.99) in Form A and 15.84 (*SD* = 7.29) in Form B for English-monolingual students. The difference between the groups was not statistically significant: $t(108) = 1.822$, $p = .071$ for Form A; $t(109) = 1.579$, $p = .117$ for Form B.

Within the typical range of 13–15 items that students reached within 3 minutes, my intention was to maximize the number of included items, so that I could evaluate DIF across as many items as possible. However, for Item 15 in Form B, all multilingual students in this study answered it incorrectly. This resulted in only one response category, posing constraints on the use of certain R packages and functions (described later) for constructing multiple-group IRT models and assessing DIF. Consequently, I decided to exclude Item 15 in Form B, using Items 1–15 in Form A and Items 1–14 for the analysis.

When assessing DIF and DTF, I used scores derived from two different scoring methods, WSC and CLS, and presented the findings separately. As mentioned previously, this decision was informed by the understanding that psychometric properties are associated with the scores and their interpretation, rather than being inherent characteristics of the instrument itself (AERA/APA/NCME, 2014; Espin & Deno, 2016). Thus, providing psychometric evidence for these two commonly used metrics separately was considered essential.

With WSC, a student's response to an item (word) was dichotomously scored, yielding two response categories (0 = incorrect, 1 = correct). Conversely, with CLS, a student's response to an item was polytomously scored. For instance, if a student writes "PAJE" instead of "PAGE," the two adjacent, correctly placed letters PA can be scored as a CLS, while AJ and JE are scored as incorrect letter sequences. Using the caret method, an upward caret (^) is placed above the correctly placed adjacent letters (P^A), and a downward caret (v) is placed below the letters for incorrectly placed adjacent letters (AvJ, JvE). CLS is also scored at the beginning and end of the word if the first and last letters are correct, respectively. Thus, "PAJE" would be scored as ^P^AvJvE^, yielding three CLS scores. In another example, if a student writes "COD"

instead of “CUT,” the response would be scored as ^CvOvDv, resulting in one CLS score. The potential range of CLS scores for each item was 0 to +1 from the number of letters within a word; for example, in a four-letter word like “PAGE”, students could attain CLS score of 0 (e.g., nothing written), 1 (e.g., vJvE^), 2 (e.g., vFvG^E^), 3 (e.g., ^PvEvG^E^), 4 (e.g., ^P^AvEvG^E^), and 5 (^P^A^G^E^), resulting in five response categories.

Of note, some students did not produce any letters for any items within each form, which included one multilingual and five English-monolingual students for Form A, and one multilingual student (the same student mentioned) and 13 English-monolingual students (including four students from Form A) for Form B. These responses were not included in the analysis.

IRT Modeling

I used IRT modeling as my primary approach for evaluating DIF and DTF. IRT is a statistical framework aimed at modeling the relation between a person’s ability (i.e., latent trait) and the probability of endorsing an item (or response category) within a given measure (de Ayala, 2022; Embreston & Reise, 2000). This method is preferred over others, such as the Mantel-Haenszel method or logistic regression (e.g., Pendergast et al., 2017), which are further described in the DIF sensitivity analysis section later in this Method section.

Researchers have developed various IRT models to analyze dichotomous and polytomous items. When dealing with WSC scores, I used the Rasch model (Rasch, 1960), whereas for CLS scores, I used the partial credit model (PCM; Masters, 1982), which is a Rasch-like model for polytomous items. Below, I provide a brief description of the Rasch model and PCM.

Rasch Model for Words Spelled Correctly Scores. Rasch model is often described as the one-parameter logistic (1PL) model because it contains one item parameter. Rasch model predicts the probability of the response of 1 (a correct response) on a dichotomous item *j* as follows (below, I used a logistic function, but other functions such as the normal ogive model can also be applied):

$$\mathrm{P}(X_j = 1|\theta, \delta_j) = \frac{exp(\theta-\delta_j)}{1+ exp(\theta-\delta_j)}$$

where θ is the person's location or latent trait level (the subscript *i*, indicating individuals, was omitted for simplicity), and δ_j is the location of item *j* or item difficulty. In simpler terms, the probability of a particular individual giving a correct response to a particular dichotomous item is determined by the distance between the individual's latent trait level and the item's difficulty level. Item difficulty quantifies the level of the latent trait required for having a 50% chance of endorsing the item; hence, a higher estimate indicates a more difficult item. Both item difficulties and latent traits are represented on the same scale (in standard deviation units) and typically range from -3 to 3, although theoretically, they span from negative to positive infinities.

The Rasch model allows items to only differ in their difficulties. As such, all items are assumed to possess equal discriminating capacity among respondents. Due to the constraint on item discrimination, item response curves—the curves that are available for each individual item and depict the relation between the latent trait and the probability of endorsing the item—are parallel across all items. This differs from more complex IRT models like the two-parameter logistic (2PL) model, as well as three- and four-parameter logistic models, where items are allowed to vary not only in their difficulties but also in other item characteristics such as discrimination, guessing, and upper asymptote.

This characteristic of the Rasch model also impacts the type or nature of DIF that can be investigated. With the Rasch model, DIF is characterized by a significant difference in the item difficulty estimates between the focal and reference groups (explained later), given the same latent trait level. This between-group difference is observed unidirectionally across all levels of the latent trait, indicating that one group consistently exhibits a higher probability of answering the item correctly than the other group across all ability levels. This type of DIF is called uniform DIF or parallel DIF. In contrast, in more complex IRT models, other types of DIF, such as non-uniform DIF (where one group shows a higher probability up to a certain level of the latent trait, after which this tendency reverses), can also be explored. Despite such limitation on the type of DIF that can be investigated, I opted for the Rasch model in this study because the existing literature suggests the use of a more parsimonious IRT model for DIF analyses when dealing with small sample sizes (Belzak, 2020; Paek & Wilson, 2011).

Partial Credit Model for Correct Letter Sequences Scores. PCM is a Rasch-like model for the polytomous data that are ordered. “Polytomous” in this context means that there are more than two response categories for individual items. “Ordered” implies that some responses reflect more or less of what is being measured compared to others, such as higher credit indicates a greater degree of correctness of the response. PCM is well-suited for analyzing item responses in achievement tests where partially correct answers are possible (e.g., multipart math problems). It is assumed that if partial credit is assigned for partially correct responses, it can offer useful information for estimating a student’s ability level.

PCM operates on the idea that each pair of adjacent categories within a polytomously scored item can be treated as dichotomous categories. Given a student’s ability level, the

probability of achieving a category score of *h* rather than a category score of *h*-1 can be obtained by applying the dichotomous Rasch model to the adjacent pair of scores. Assuming item *j* is scored $x = 0, 1, \ldots, m_j$ with K_j response categories ($K_j = m_j + 1$). The probability that a person obtains a category score of X_j is given by the formula:

$$\mathrm{P}(X_j \mid \theta, \delta_{jh}) = \frac{exp\left[\sum_{h=0}^{x_j}(\theta - \delta_{jh})\right]}{\sum_{k=0}^{m_j} exp\left[\sum_{h=0}^{k}(\theta - \delta_{jh})\right]}$$

where $\sum_{h=0}^{0}(\theta - \delta_{jh}) \equiv 0$. Again, θ represents the person's latent trait level. δ_{jh} (with $h = 1, \ldots, m_j$) refers to the transition location parameter for item *j* associated with a category score of *h*, which is often called item threshold or item step difficulty. As indicated by the subscript *j* on *m* (m_j), the number of category scores can vary across items.

Just like the Rasch model, PCM solely focuses on item difficulty. However, in the case of polytomous items having multiple response categories, there are multiple item thresholds (δ_{jh}) within an item, each representing the relative difficulty in endorsing category *h* over category (*h*-1). Similar to item response curves in the Rasch model, there are category response curves (i.e., option response functions, option characteristic curves, operating characteristic curves) for polytomous items, graphically illustrating the probability of achieving a specific category score as a function of the latent trait.

PCM serves as a useful model for analyzing CLS scores within Word Dictation. When an item is scored for CLS, partial credit (e.g., assigning 1 CLS) is allocated for correctly writing two adjacent letters in order within a given word. For instance, using the previously mentioned example item "CUT" with five response categories, four thresholds or step difficulties can be posited: δ_{j1}(0 CLS vs. 1 CLS), δ_{j2} (1 CLS vs. 2 CLS), δ_{j3} (2 CLS vs. 3 CLS), δ_{j4} (3 CLS vs. 4

CLS). Of important note, in the case of Word Dictation CLS scores, the number of response categories varies across items; for instance, while an item "CUT" has five category scores, another item "PAGE" has six category scores.

Since polytomous items consist of multiple response categories, detecting DIF for this item type requires comparing multiple thresholds or step difficulties within an item simultaneously across the focal and reference groups (Baker, 1992). There are generally two classes of omnibus polytomous DIF statistics: net DIF statistics and global DIF statistics. Net polytomous DIF statistics evaluate the aggregated net effect across all category scores of the polytomous items. Thus, a difference in the sign of the DIF effect across category score levels (e.g., favoring one group for some category score levels but favoring the other group for other score levels) can result in a net DIF effect of zero (or near zero), despite sizable effects within particular category score levels. In contrast, global DIF statistics assess the unsigned effect across all category scores within the item; hence, this approach is generally considered sensitive to DIF impacts with different signs across category score levels (Penfield, 2007).

As described later, my primary approach for evaluating DIF is the IRT-based likelihood ratio test, and this method uses global DIF statistics for polytomous items. Additionally, I also use Raju's DFIT approach as a supplementary method to complement the likelihood ratio test, and this approach adopts net DIF statistics. Of note, neither global nor net DIF statistics provide information regarding precisely which category score levels are responsible for the observed DIF effect (Penfield et al., 2008). In other words, if DIF is detected for a specific Word Dictation item scored with CLS, this study lacks the capacity to pinpoint which score categories contribute to the DIF detected.

Evaluating Assumptions of Item Response Theory Models

When using the IRT framework, two fundamental assumptions are typically required: *local independence* and *unidimensionality*. In essence, local independence, also called conditional independence, requires that all pairs of item responses are independent, given the latent trait level (Embreston & Reise, 2000). That is, the probability of endorsing an item is determined solely by the students' trait level, rather than their responses to other items. The unidimensionality assumption means that the responses to the items are solely a function of a single continuous latent trait variable (de Ayala, 2022). That is, a single latent construct is adequate to account for the shared variance among item responses. The violation of local independence can be considered an indication of a potential violation of unidimensionality, as local dependence often occurs when one or more additional dimensions, beyond the intended dimension, influence the item response (Embretson & Reise, 2000; McDonald, 1981).

Local Independence. I used Q_3 (Yen, 1984) and the Jackknife Slope Index (JSI; Edwards et al., 2018; Houts & Edwards, 2013) statistics to assess local independence. The Q_3 and JSI statistics allow for identifying pairs of items that exhibit interdependence.

The Q_3 statistic denotes the correlation between residuals—the difference between observed and model-expected item responses—of any two items across all the respondents who answered those two items. This correlational statistic ranges between -1 and +1. The absolute value of 1 indicates that the two items are perfectly dependent.

Similarly, the JSI index also captures the interdependence between two items. JSI values represent the slope change in the *j*th item resulting from the removal of the *k*th item, as well as the slope change in the *k*th item resulting from the removal of the *j*th item. In formula, $JSI_{j(k)} =$

$\frac{a_j - a_{j(k)}}{se(a_j)}$ where a_j represents the full data slope estimate, $a_{j(k)}$ represents the slope estimate of item *j* with item *k* removed from the dataset, and $se(a_j)$ represents the standard error of the slope parameter when all items are considered.

To obtain these statistics, I used the *LOCALDEP* function in *EFA.dimensions* package (O'Connor, 2023). Regarding the cutoff value, I applied the commonly used criterion of $|Q_3| \geq 0.2$ as indicative of the potential violation of local independence (D. Kim et al., 2011). This cutoff value was initially suggested by Yen (1993), particularly in the context of measures with at least 17 items (de Ayala, 2022); although the number of items in this study is slightly smaller (14–15), I deemed it acceptable to apply this criterion.

Regarding the JSI values, an ad hoc criterion was calculated as recommended in the literature (Houts & Edwards, 2013). Obtained JSI values are aggregated, resulting in a lower triangular matrix with $(k) \times (k-1)$ elements. An ad hoc criterion is then determined as the mean of these lower triangular elements plus 2 times the standard deviation. JSI values exceeding this criterion value are flagged as locally dependent. Simulation studies have shown that this index can effectively detect local dependence even with a relatively small sample size (e.g., *N* = 250; Edwards et al., 2018), as in this study.

Unidimensionality. I used several techniques to check the unidimensionality assumption. First, for the Rasch model with WSC scores, I used the modified parallel analysis approach proposed by Drasgow and Lissak (1983). Specifically, I compared the eigenvalues obtained from a factor analysis of the observed dataset with those generated under the assumed unidimensional IRT model. The null hypothesis of unidimensionality is rejected if the second eigenvalue is substantially larger for the observed dataset compared to that from the simulated data. I used 100

simulated samples via the Monte Carlo procedure to approximate the distribution of the statistic under the null hypothesis. This analysis was performed using the *unidimTest* function in the *ltm* package (Rizopoulos, 2006).

Given that the *unidimTest* function currently does not support the analysis with polytomous data, I used an alternative approach to assess the unidimensionality of PCM with CLS scores. I conducted exploratory factor analysis (EFA) (Asparouhov & Muthén, 2009), specifically by performing a principal component analysis (PCA) based on singular value decomposition, using mean-centered scores, and looking at the scree plot. For this purpose, I used the *prcomp* function embedded in R.

Evaluating Model Fit

To evaluate the model fit of the unidimensional Rasch model and PCM, I used the M_2 (Maydeu-Olivares & Joe, 2005, 2006) statistic as an absolute model fit measure for dichotomous data, and the collapsed M_2* statistic (Cai & Hansen, 2013) for polytomous data. The M_2 statistic is part of the limited information goodness-of-fit statistics family and tests the null hypothesis stating that the observed response data is not statistically significantly different from the modeled response data. This statistic is considered superior for evaluating the goodness of fit for IRT models compared to full-information absolute model-data fit tests, such as Pearson's chi-square and its variations, by addressing the sparse data issue (Joe & Maydeu-Olivares, 2010; Maydeu-Olivares & Garcia-Forero, 2010). Additionally, I included other fit indices serving as continuous measures of model-data correspondence. These indices include: root mean square error of approximation (RMSEA; Steiger, 1990), standardized root mean square residuals (SRMSR; Jöreskog & Sörbom, 1981), comparative fit index (CFI; Bentler, 1990), and Tucker-Lewis Index

(TLI; Tucker & Lewis, 1973). I interpreted CFI ≥ .95, TLI ≥ .95, SRMSR ≤ .08, and confidence intervals for RMSEA close to 0 as evidence of a good model fit, thereby supporting the assumption of unidimensionality.

Evaluating Differential Item Functioning for Individual Items

As my primary method for detecting DIF, I used the IRT-based likelihood ratio model comparison approach (Thissen et al., 1993). To perform this analysis, I used *multipleGroup* and *mirt.model* functions in the *mirt* package (Chalmers, 2012) to construct multiple-group IRT models, using a full-information maximum likelihood estimation. Subsequently, I used the *DIF* function within the same package to conduct DIF analyses. Specific procedures are described in detail below.

Reference and Focal Groups. I designated English-monolingual students as a reference group and multilingual students as a focal group. In the context of DIF detection, the reference group is typically identified as the majority sociodemographic group with a larger sample size, whereas the focal group represents the minority sociodemographic group with a smaller sample size. Such a designation is arbitrary and does not affect the computation of DIF.

Identifying Anchor Items. In DIF analyses, it is typical to include anchor items, allowing for placing all items in the focal and reference groups on a common metric (Tay et al., 2015). It is essential to use appropriate anchor items, as including items with DIF as anchor items can impact the overall DIF results and decisions (e.g., Jodoin & Gierl, 2001). A common approach is to use a set of items known as measurement invariant or DIF-free (i.e., known-anchor approach). However, in the current study, there were no predetermined DIF-free items from previous investigations. Hence, for dichotomous DIF analysis, I used a three-step procedure

following an approach outlined by Tay et al. (2015). This approach initially focuses on empirically selecting anchor items, followed by DIF detection.

First, I began by constructing multiple-group IRT models where all item parameters were constrained to be equal across both groups. This allowed for freely estimating the latent mean and variance of the focal group relative to the reference group. Second, I established preliminary linking between the two groups by treating the estimated latent mean and variance as fixed. In this step, all the item parameters were freely estimated and preliminarily tested for DIF. This step enabled the empirical selection of DIF-free items. Lastly, using these selected items as anchors, the remaining items were tested for DIF.

Subsequently, in polytomous DIF analysis with CLS scores, I used the items identified as non-DIF in WSC (i.e., those not flagged as having DIF in WSC, based on the likelihood ratio test and all other supplemental approaches, as will be described later) as anchor items, following the known-anchor approach. This approach assumes that items without DIF in WSC scoring will also lack DIF in CLS. While this assumption was deemed reasonable for the current analysis, future studies with CLS scoring may require a more nuanced approach. This point is further discussed in the Recommendations for Future Research section.

Likelihood Ratio Model Comparison. In essence, this method tests whether the less constrained model (which allows for DIF on some items) differs significantly from the more constrained model (which assumes no DIF and constrains the item parameter to be identical across groups). Specifically, the constrained (no DIF) model assumes that the subgroups being compared respond to each item in the same way, thereby fixing the item parameters (such as item difficulty) to be identical for all groups. In contrast, the less constrained (DIF) model allows

for the possibility that subgroups respond differently to some items, relaxing the constraints and permitting item difficulty to vary between groups. The likelihood ratio test is used to determine whether the difference in fit between these two models is statistically significant. It compares how well each model fits the data, and if the less constrained model fits significantly better, it suggests the presence of DIF. In the current analysis, likelihood ratio chi-square statistics with *p*-values smaller than 0.05 were considered flagging DIF. I used the adjusted *p*-values based on Benjamini-Hochberg's procedure (Lord, 1980; Marco, 1977) to mitigate inflated Type 1 error rates due to multiple comparisons. The Benjamini-Hochberg method (alongside the Holm method) is deemed optimal for DIF detection purposes (Kim & Oshima, 2013).

Effect Size Estimates. The power of statistical significance tests, however, is influenced by sample size. With a small sample size like in this study, it is possible that the null hypothesis of no DIF will not be rejected even if there is a large DIF effect size. Additionally, a statistically significant difference in item parameters does not necessarily imply a practically meaningful difference. Thus, I evaluated the magnitude or severity of DIF using DIF effect size indices.

Various conceptualizations and reporting methods exist for DIF effect sizes (Chalmers, 2023; DeMars, 2011). In this study, I used the Expected Score Standardized Difference (ESSD), which measures the differences in expected scores (i.e., expected item score for a latent trait level given the set of estimated item parameters) between the focal and reference groups (Meade, 2010). Following Cohen's *d* interpretation, this metric provides a standardized measure of difference and is applicable to both dichotomous and polytomous item DIF. I used the *empirical_ES* function in the *mirt* package to obtain the effect size estimates. Note that I derived the DIF effect sizes from the final multiple-group IRT model, where items not identified as

having DIF (collectively considering the results of all DIF detection methods) were treated as anchors and held fixed as equal across groups.

Evaluating Differential Test Functioning for Word Dictation as a Whole

Following the DIF analysis, I assessed DTF at the assessment level. Assessing DTF is valuable because it allows for determining whether Word Dictation, as a whole—the first 15 items in Form A and 14 items in Form B—, functions differently for English-monolingual and multilingual student groups. Aligning with the effect size metrics used for DIF, I reported the Expected Test Score Standardized Difference (ETSSD), which is Cohen's *d* for expected test scores. These values were again computed using the *empirical_ES* function.

Differential Item Functioning Sensitivity Analyses

DIF is complex in nature, and no single detection strategy suffices for a comprehensive examination interpretation. To supplement the likelihood ratio test results, I conducted additional analyses employing different statistical frameworks (e.g., IRT-based vs. observed score-based) and DIF detection techniques (e.g., comparing item parameters between groups vs. examining area between item response functions).

In deciding whether an item exhibits DIF or not, my interpretation primarily relied on the IRT-based likelihood ratio test, given its effectiveness in detecting DIF when dealing with small sample sizes (Agresti, 2002; Belzak, 2020). However, for thoroughness, I also considered items identified as showing DIF by any of the supplemental methods outlined below. For the items identified as having DIF by at least one method, I generated item response functions to visually represent the direction and magnitude of DIF and reported the DIF effect sizes. When assessing

DTF at the assessment level, I allowed these flagged items to be freely estimated, ensuring that all their effects were taken into consideration.

For dichotomous WSC scores, the additional approaches included Mantel-Haenszel's method (Holland & Thayer, 1988), logistic regression (Swaminathan & Rogers, 1990), Raju's area method (Raju, 1988, 1990), and Raju's DFIT framework (Raju et al., 1995). Mantel-Haenszel's method and logistic regression are based on observed scores (CTT framework). Raju's area and DFIT methods, in contrast, are rooted in IRT modeling. Instead of focusing on the differences in item parameters between the two groups, as done by the IRT likelihood ratio test, these methods emphasize the examination of how item response functions of the two groups vary. In essence, Raju's area method evaluates whether this gap is statistically significant. Raju's DFIT framework provides various indices to quantify the magnitude of such gaps, including non-compensatory differential item functioning (NCDIF), with a particular focus on how the results of the focal group are affected by DIF, if present (Cervantes, 2017).

I used the *difR* package (Magis et al., 2020) for Raju's area method, Mantel-Haenszel's method, and logistic regression; this package is specifically designed for dichotomous data, currently not supporting polytomous item DIF analysis. In these analyses, an item purification process—iteratively removing items identified as exhibiting DIF from the set of anchor items, aiming to enhance the accuracy of DIF detection (Candell & Drasgow, 1988)—was implemented. I used adjusted *p*-values based on the Benjamini-Hochberg method to account for multiple comparisons.

For the DFIT method, I used the *DFIT* package (Cervantes, 2017). The *DFIT* package supports polytomous item DIF analysis as well. Given the relatively limited availability of

packages and functions for assessing DIF with polytomous items, especially those with varying numbers of response categories, I focused on reporting likelihood ratio test and DFIT indices for polytomous item DIF analysis with CLS scores. Among various DFIT indices, I examined NCDIF values.

Researchers proposed cutoff values of NCDIF > .006 for dichotomous items and NCDIF > $.006(k-1)^2$ for polytomous items, with *k* representing the number of response categories, based on simulation studies (Oshima & Morris, 2008; Raju et al., 1995). However, these cutoffs do not consider the unbalanced sample size, which is pertinent in this study. One advantage of using the DFIT package is its capability to compute cutoff points for various indices, including NCDIF, through the Monte Carlo item parameter replication (IPR) procedure. Essentially, cutoff values can be derived directly from information on the estimated item parameters and sample sizes of the groups, where item parameters from the focal group serve as the expected values for both groups, but the asymptotic variance-covariance matrices are calculated for each group based on their respective sample sizes and latent ability distributions (see Cervantes, 2017 for details).

To perform this, I used the *AseIrt* and *CutoffIpr* functions in the *DFIT* package. In terms of the item parameters, I used those derived from the multi-group IRT models established via the *mirt* package. Using the Monte Carlo IPR with 1000 replications, I obtained cutoff points for each individual item and considered NCDIF values above them as indicative of uniform DIF. I also analyzed post-hoc power for NCDIF values using *Ipr* and *IprNcdif* functions, indicating the power to detect the true NCDIF considering the sample sizes of each group.

The package does not currently support this approach to obtaining cutoff points for polytomous items. Thus, for polytomous items, I used the cutoff criteria based on existing

literature (Oshima & Morris, 2008; Raju et al., 1995). For example, for a polytomously scored item with five response categories like “CUT,” I used the adjusted criterion of NCDIF > .096.

Chapter 4: Results

In this chapter, I present the findings of this dissertation, which investigates the extent to which the CBM Word Dictation demonstrates measurement invariance concerning the language status of students with intensive writing needs. First, I begin by outlining the results of preliminary analyses, including descriptive statistics of Word Dictation scores, examination of data missingness patterns, and checks of assumptions for IRT modeling. Next, I report the findings for Research Question 1, which focuses on item-level measurement invariance for WSC and CLS scores. Last, I present the findings for Research Question 2, focusing on the full assessment level.

Preliminary Analyses

Descriptive Statistics for Word Dictation Scores

Table 10 presents descriptive statistics for students' Word Dictation total scores, encompassing 30 items within each form. The table includes scores based on different scoring metrics (words written, WSC, CLS). Overall, scores for words written and CLS generally approximated normal distribution, but WSC scores were positively skewed with a sharp peak and heavy tails. There was no statistically significant difference in assessment-level scores between multilingual and English-monolingual student groups, with Hedges' *g*s of 0.

Table 11 presents descriptive statistics of item-level scores, focusing on WSC and CLS. Across the entire sample, for WSC scoring in Form A, item-level mean scores ranged from .08 ("SWAP") to .52 ("PET"), indicating that between 8% and 52% of students in the current sample answered each item correctly. In Form B, WSC mean scores ranged from .04 ("MOAN") to .49 ("LIP"). In the current sample, even the item with the highest proportion of student endorsements

reached only around 50% accuracy when scored with WSC. When scored with CLS, item means ranged from 1.96 (“SWAP”) to 2.88 (“PAT”) for Form A and from 1.63 (“KEPT”) to 2.77 (“PROP”, “LIFT”, “CROP”) for Form B. The table also includes item-total correlation coefficients (Pearson’s *r*s), overall suggesting good discrimination (within the CTT framework).

Data Missingness Pattern

A nontrivial proportion of missing responses (i.e., student not writing any letters for a given item) was evident within the initial 15 (Form A) or 14 (Form B) items (see Table 11), although not as substantial as the missingness observed in later items. To understand the missingness pattern within this subset of items (and to take any action in the consequent analyses if needed), I conducted a missingness analysis using the *finalfit* package (Harrison, 2020).

I examined whether the data loss pattern correlated with other variables, specifically sex, grade, special education eligibility, and language status. Three general data loss mechanisms are available (Rubin, 1976): *missing completely at random* (missingness in a variable is unrelated to all other variables in the observed dataset and unrelated to the variable itself with no missing data), *missing at random* (missingness is correlated with other variables in the dataset but is independent of the level of the variable itself with no missing data), and *missing not at random* (missingness is linked to the true level of the variable, not observed in the dataset).

Table 12 presents the chi-square test results for Item 15 and Item 14 for Forms A and B, the two items having the highest proportion of missing data among the items analyzed. The missing data did not show statistically significant associations with students’ sex (*p*s = .510 and .648) and special education eligibility (*p*s = .855 and .939). However, missingness was statistically significantly associated with students’ grade levels (*p*s < .001) and language status

(*p*s = .025 and .043). Specifically, a significantly higher percentage of missing responses was noted among students in lower grade levels compared to those in higher grade levels; this pattern emerged not only in these two items but more generally, starting from relatively early items (the 4th or 5th item in each form). In terms of language status, a higher proportion of multilingual students had missing responses compared to English-monolingual students; this pattern was observed in other items as well, particularly in relatively later items (Items 14 in Form A and Items 10–11 in Form B).

Taking these results into consideration, I concluded that the data loss pattern is *missing at random*, as the factors influencing the occurrence of missing data can be accounted for within the observed data. In other words, considering the grade levels and language status, I can reasonably assume that the missingness in the items occurred randomly. Under the missing at random mechanism, the full information maximum likelihood estimation approach, which I used in the subsequent IRT model analyses, generally yields unbiased estimates by effectively handling the missing data issue.

Item Response Theory Model Assumptions

Local Independence. Tables 13 and 14 present the Q_3 and JSI values for the 105 pairs of items in Form A and 91 pairs of items in Form B, scored based on WSC scores. Tables 15 and 16 provide the Q_3 and JSI values for the same sets of item pairs, this time scored using CLS scores. Note that these values were reported for the full sample.

Overall, although several pairs showed signs of local dependence, the number of such pairs was minimal. With WSC scores, 6.7% based on Q_3 values and 2.9% based on JSI for Form A, and 8.8% based on Q_3 values and 3.3% based on JSI for Form B indicated potential local

dependence. With CLS scores, 5.7% based on Q_3 values and 1.9% based on JSI values indicated local dependence for Form A, while 2.2% based on Q3 and 3.3% based on JSI values for Form B. Thus, I considered it reasonable to assert that the local independence assumption was supported, further reinforcing confidence in the validity of the unidimensionality assumption described below.

Unidimensionality. For the dichotomous Rasch model with WSC scores, the modified parallel analysis results did not reject the null hypothesis of unidimensionality. For Form A, T_{obs} (i.e., the second eigenvalue for the original dataset) = 1.163, T_b (i.e., the average of second eigenvalues in the Monte Carlo samples) = 1.109, and p = .327. For Form B, T_{obs} = 1.465, T_b = 1.316, and p = .208. Scree plots (Figure 2) for eigenvalues in observed and simulated datasets further support the validity of the unidimensionality assumption.

Concerning the PCM model with CLS scores, the unidimensionality assumption was also supported. Table 17 displays the PCA results, and Figure 3 provides scree plots of the eigenvalues of the principal components. For Form A, the first factor had an eigenvalue of 19.528, explaining approximately 55.8% of the variability, whereas the second factor had an eigenvalue of 1.952, explaining only 5.8% of the variability. For Form B, the first factor had an eigenvalue of 23.057, explaining about 60.7% variability, while the second factor had an eigenvalue of 2.020, explaining 5.3% variability. These results indicate that the first principal component explains the majority of the variation (literature suggests the variance accounted for by the first factor should ideally be larger than 20%; Reckase, 1979), further supported by the scree plots.

Lastly, when fitting the unidimensional Rasch model to the observed WSC data and unidimensional PCM to the CLS data, the model fit (Table 18) was generally deemed acceptable, despite some inconsistencies across model fit indices. In the Rasch model, M_2 statistics were statistically significant in both forms, indicating poor fit, and SRMSR values did not meet the criteria for good fit either. However, CFI, TLI, and RMSEA values were within an acceptable range or indicated good fit in both forms. In the PCM model, M_2^* statistic was statistically significant in Form A (indicating poor fit) but not for Form B. Other fit indices provided mixed results. CFI and TLI values indicated poor fit for Form A but approximated good fit for Form B. RMSEA values approximated an acceptable fit in both forms, but SRMSR did not suggest good fit in either form. Considering all of the evidence, I cautiously concluded that it is reasonable to assume unidimensionality for both WSC and CLS data.

Research Question 1: To What Extent Does Word Dictation Show Measurement Invariance at the Item Level When Analyzing Words Spelled Correctly and Correct Letter Sequences?

Differential Item Functioning Results Based on Words Spelled Correctly Scores

The results of DIF analyses using WSC scores are presented in Table 19 (left side). Using the IRT-based likelihood ratio method with adjusted *p*-value, none of the items in Form A were identified as exhibiting DIFs. However, upon considering additional supplemental methods, Item 4 (“PAT”) was found to demonstrate uniform DIF according to two methods—Raju’s DFIT statistic (NCDIF = .031) and Raju’s area method (p = .058; marginally significant). Furthermore, Items 3 (“SACK”), 10 (“SILK”), and 15 (“KIND”) were also flagged as uniform DIF based on Raju’s NCDIF statistic.

Figure 4 displays the item response functions for individual items, with separate curves drawn for the DIF-flagged items for multilingual and English-monolingual student groups. This visualization provides insight into the direction and severity of the detected DIF. The plots reveal that the direction of DIF effects varies. Specifically, for Items 10 (“SILK”) and 15 (“KIND”), multilingual students were more likely to get the items correct than English-monolingual students across all latent trait levels. However, for Items 3 (“SACK”) and 4 (“PAT”), this pattern was reversed, with English-monolingual students more likely to get the items correct than multilingual students.

Regarding the severity or magnitude of DIF, the differences in item response functions between the groups were relatively minor for Items 10 and 15 but appeared more pronounced for Items 3 and 4. This observation is supported by small effect sizes for Items 3, 10, and 15, with absolute values of ESSD ranging from .325 to .417, whereas Item 4, identified as having DIF by two methods, approached a medium-sized DIF effect (ESSD = -.498), favoring English-monolingual students.

For Form B, the results of all analysis methods indicated that none of the items analyzed showed uniform DIF. As a result, each item generated a single item response curve for both groups (see Figure 5).

Differential Item Functioning Results Based on Correct Letter Sequences Scores

Table 19 (right side) encompasses DIF analysis results for CLS scores. Note that for the analyses with CLS scores, I used the non-DIF items based on previous analyses with WSC scores as anchor items (i.e., constraining these items to be equal across the groups). As such, DIF results are not available for these anchor items and are only available for items flagged as

potentially exhibiting DIF in their WSC scores. Again, an implicit assumption in this approach to selecting anchor items is that items without DIF in WSC scoring would also not exhibit DIF in CLS. While this decision was considered reasonable for the current analysis, a more nuanced approach may be necessary in future studies with CLS scoring. I discuss this point in greater detail in the Recommendations for Future Research section of the Discussion.

In Form A, none of the items flagged as potentially exhibiting DIF using WSC scoring (Items 3, 4, 10, and 15) were found to exhibit DIF in their CLS scores according to the likelihood ratio test using PCM. However, Raju's NCDIF statistics flagged Items 4 ("PAT") and 15 ("KIND") as demonstrating uniform DIF. The item response function (Figure 6) shows that for Item 4, English-monolingual students were more likely to achieve higher CLS scores than multilingual students across all latent trait levels, though the discrepancy between the curves was relatively minor, as further supported by the small effect size of DIF (ESSD = -.399). Regarding Item 15, the item response functions, which intersect, reveal that up to a certain point of latent ability level (around 1 theta score), English-only students were more likely to receive higher CLS scores than multilingual students; however, beyond that level (for students with higher latent ability than about 1 theta score), the pattern reversed, with multilingual students obtaining higher scores than their English-monolingual students given the same ability level. Nevertheless, the item response functions were nearly overlapping, as evidenced by the near-zero effect size of DIF (ESSD = -.020).

For Form B, since no items were identified as exhibiting DIF based on the dichotomous item DIF analysis, none were evaluated for DIF. Figure 7 visually shows that all items have a

single item response curve, rather than showing two separate curves for multilingual and English-monolingual student groups.

Research Question 2: Does Measurement Invariance Hold at the Assessment Level?

After identifying item-specific DIF, I assessed DTF using the final multiple-group IRT model, where used where items not identified as having DIF served as anchors and were held constant across groups. Analysis of WSC scores indicated that Form A did not exhibit notable DTF between the two groups. The effect size of DTF was minimal (ETSSD = -.017). For Form B, where no DIF items were found, the DTF effect size was computed as 0, indicating the absence of DTF.

Examining DTF results based on CLS scores, the effect size for Form A was also negligible (ETSSD = -.027), suggesting measurement invariance at the assessment level across the groups. For Form B, mirroring the findings from WSC scores, the DTF effect size was again calculated as 0 since all items were used as anchors and held constant across groups. Note that since I only examined the subset of Word Dictation items, these DTF results may differ when students complete all 30 items or when responses to additional items are considered.

Chapter 5: Discussion

The goal of this dissertation was to evaluate the measurement invariance of CBM Word Dictation, designed to assess English transcription (spelling and handwriting) skills at the word level, for multilingual students and English-monolingual students identified as having intensive early writing needs. I identified several items, especially in Form A, as potentially displaying DIF based on WSC and CLS scores. When comparing the scoring metrics, some differential item functioning observed in words spelled correctly was mitigated when using correct letter sequences. The direction of such DIFs was not consistent across the detected items, with some items favoring the multilingual student group over the English-monolingual student group, while others showed the opposite pattern. The magnitude of detected DIFs was overall small. When it comes to the assessment level, neither Form A nor Form B exhibited differential functioning across the two student groups. Below, I elaborate on the implications of these findings, the limitations of the study, and directions for future research and practice.

Some Items Functioned Differently Across the Groups

I did not hold a strong hypothesis regarding whether and which specific items would function differently for multilingual and English-monolingual students with the same level of English transcription skills. However, I did expect that some of the DIF detected with WSC may be mitigated under the CLS scoring method.

Based on WSC scores, a few items in Form A did emerge as potentially exhibiting DIF, specifically the words “SACK,” “PAT,” “SILK,” and “KIND.” The direction of DIF varied among these identified items; while multilingual students were less likely to correctly write “SACK” and “PAT” compared to English-monolingual students with the matched ability levels,

they were more likely to write “SILK” and “KIND” correctly. The effect sizes of DIF were mostly small, although a close to medium-sized DIF was observed for “PAT.” However, as there has been no prior evaluation of DIF for this measure, it is essential for future research to replicate this study with a similar population.

Some Differential Item Functioning Identified with Words Spelled Correctly were Alleviated with Correct Letter Sequences

Among items identified as exhibiting DIF based on WSC scores, some did not show statistically significant DIF when using CLS scores (“SACK” and “SILK”), or if still present (“KIND”), the magnitude of DIF approached zero. This indicates that when these items were scored using CLS, multilingual students were no more likely to score lower on the word “SACK” due to their group membership, and similarly, English-monolingual students were no more likely to score lower on the words “SILK” and “KIND.” These findings tentatively suggest that when examining students’ item-level performance, CLS may provide more valid information than WSC, at least for the aforementioned items.

This variability in item-level measurement invariance depending on the specific scores is not surprising. Previous research on English writing CBMs has provided differing levels of evidence for validity depending on the scoring method (e.g., Romig et al., 2017, 2021) and thus highlighted the importance of tying psychometric characteristics of tasks to specific scores, rather than the tasks themselves (Espin & Deno, 2016). Furthermore, CLS, in particular, has been recognized as a valuable scoring metric for early writing assessment, sensitively capturing writing growth over time (Choi et al., 2023; Hampton & Lembke, 2016) and showing robust correlations with standardized, norm-referenced writing measures at given time points (Hampton

& Lembke, 2016; Lembke et al., 2003). Moreover, findings from my literature review in Chapter 2 focusing on the multilingual student population also supported CLS; the reviewed studies provided stronger criterion validity evidence when using CLS (*r*s = .56–.78) compared to other metrics like WSC (*r*s = -.00–.81) and words written (*r*s = -.05–.66), across different tasks. The current finding adds to the existing knowledge about the technical quality of CLS scoring, particularly emphasizing its quality *at the item level*, with minimal levels of DIF observed (both in terms of quantity and severity).

Exploring Cross-Linguistic Transfer as a Possible Influence on Observed Differential Item Functioning

This statistical detection of DIF should not be interpreted as indicating item *bias.* To make any inference regarding bias, further research is needed, such as conducting item reviews by content experts, to confirm whether the differences in item responses between the groups are indeed attributable to a construct-irrelevant factor (here, language status) systematically. Although delving into the underlying causes of the DIF falls beyond the scope of this study, I outline some possible scenarios to provide preliminary insights.

When assessing multilingual students, measurement variance may arise from various factors, including students' familiarity with response formats, characteristics of assessments or items, examiners' testing styles or levels of specialized training (AERA/APA/NCME, 2014; Sandberg & Reschly, 2011). Within this study context, it is less likely that the response format and a certain student group's lesser familiarity with it are the primary causes of DIF. Word Dictation represents a type of activity that teachers frequently use in their classrooms, and all

Word Dictation items share identical response formats. Any student group still less familiar with the response format would have typically benefited from the practice item.

Alternatively, one possible explanation is the occurrence of cross-linguistic transfer (Chung et al., 2019; Cummins, 1979; Koda, 2008; Lado, 1957; MacWhinney, 2005) when multilingual students attempted to answer certain items, especially those identified as exhibiting DIF. This situation might have arisen if those items contained cues or shared features across English and the student's home language (Cummins, 1979; MacWhinney, 2005). If cross-linguistic transfer did indeed occur, it could have influenced how multilingual students responded to the items in two different ways: either enhancing English spelling accuracy (indicating DIF favoring the multilingual student group) or impeding it (indicating DIF favoring the English-monolingual student group).

The former scenario might have arisen if the multilingual student possessed grapheme-phoneme knowledge in their home language and if the given Word Dictation items involved letters with shared grapheme-phoneme correspondences between the home language and English. This could prompt the student to source their home language when spelling the English word. Alternatively, the latter situation could have occurred if the specific phoneme involved in the given word is absent in the student's home language but rather prompts nearby phonemes available. This could cause confusion and lead to English spelling errors (e.g., Spanish-English bilingual students often write *fonny* for *funny*, as mentioned in the Introduction; Bahr et al., 2015; Wolters & Kim, 2024). However, it remains unclear and requires further investigation whether the Word Dictation items "SILK" and "KIND" fit into the first case and whether the items "PAT" and "SACK" fit into the second case.

It is important to note that these discussions regarding cross-linguistic influence should only be considered as hypotheses and interpreted with caution. Considerations of cross-linguistic influence should be situated within specific L1 contexts. However, the multilingual student group in the current DIF analysis represented a diverse range of L1s. Furthermore, to speculate on whether a cross-linguistic transfer occurs, it is essential to understand the extent to which the multilingual student has developed grapheme-phoneme knowledge in their L1, along with the type of language support the student has received. This detailed information was unavailable in this study. As a result, it remains unclear whether the multilingual students in this study were actually able to source their L1 when responding to Word Dictation items.

Disability*: *Ignored Aspect in the Current Differential Item Functioning Analysis

An implicit assumption behind the current DIF analysis is that any disparities observed in item responses between the multilingual and English-monolingual student groups are solely attributable to their language status. However, this classification based on language status could have actually served as a proxy for other differences between the groups (Tay et al., 2013). One such factor is disability characteristics.

As outlined in the Method section, a larger proportion (93.3%) of English-monolingual students were receiving special education services compared to multilingual students (64.2%). Consequently, the current analysis might involve situations where some multilingual students primarily struggling with language acquisition and not eligible for special education services were compared with English-monolingual students having language delays or disabilities. In this scenario, disability may have played a role in the observed language differences in students' item responses. Moving forward, it would be advantageous to simultaneously consider disability

characteristics alongside language factors, either aiming to control the impact of disability and focus on language differences or to examine both as potential causes of DIF. To achieve this, researchers may find it beneficial to use the multiple indicators multiple causes (MIMIC) model within a confirmatory factor analytic framework (Muthén, 1985, 1988) and the IRT with covariates (IRT-C) model (Tay et al., 2011).

Word Dictation Did Not Function Differently Across the Groups at the Assessment Level

I hypothesized that Word Dictation, in its entirety, would not demonstrate measurement invariance concerning students' language status, which was confirmed in this study. Findings suggest that students with similar levels of English transcription skills were expected to achieve similar scores in Word Dictation, regardless of whether they were categorized as multilingual or English-monolingual students. This finding remained consistent regardless of which scoring metrics (WSC and CLS) were used. This lack of DTF is indeed not unexpected in light of the DIF findings; some items favored multilingual students, while others favored English-monolingual students, canceling out the impact across the assessment level.

This confirmation of assessment-level measurement invariance for both WSC and CLS scores adds empirical evidence supporting the validity and fairness of using Word Dictation for linguistically diverse students struggling with English early writing. While an extensive body of research has established the technical rigor of various English writing CBMs (McMaster & Espin, 2007; McMaster et al., 2011b; Romig et al., 2017, 2021; Shin & McMaster, 2019), including Word Dictation (Choi et al., 2023; Hampton & Lembke, 2016; Lembke et al., 2003), limited information has been available for multilingual students (Keller-Margulis et al., 2016; Sandberg & Reschly, 2011; Smith & Lembke, 2022). A smaller body of literature has begun to

address this gap (Smith et al., 2023), and the findings of this study contribute to the growing body of research in this area.

This study further expands the current knowledge in a unique way. First, to the best of my knowledge, this study represents the first effort within the field of CBM English writing to empirically evaluate the measurement invariance of the task, with a clear focus on issues of fairness and structural validity (AERA/APA/NCME, 2014; Messick, 1995; Truckenmiller et al., 2022). Ensuring measurement invariance is imperative to ensure that the measure captures the intended construct equivalently across subgroups, enabling meaningful interpretations of the observed scores for each group (Raju et al., 2002). Furthermore, the finding that the measurement invariance was confirmed for Word Dictation sets the stage for any future research that involves comparisons of mean scores using this measure between multilingual and English-monolingual student groups.

Second, as a part of assessing measurement invariance, this study provides empirical support for the unidimensionality of Word Dictation based on both CLS and WSC scores for the entire sample. This finding contributes to addressing a gap in research on writing CBM examining the dimensionality of different tasks and scoring metrics (Kim et al., 2015). However, at the same time, I should highlight that not all model fit indices supported the unidimensional structure (although generally acceptable for proceeding with IRT modeling). Additionally, some item pairs showed local dependence, which is considered indicative of the presence of additional latent factor(s) (Embretson & Reise, 2000; McDonald, 1981). Thus, additional investigation into modeling the dimensionality of Word Dictation item response data is needed.

Beyond this statistical rationale, certain attributes of Word Dictation also underscore the advantages of intricately modeling the dimensionality of the measure. First, Word Dictation items are not fully equivalent to each other. Items include different word types, such as CVC (consonant-vowel-consonant) and CVCC (consonant-vowel-consonant-consonant), varying in complexity. Second, Word Dictation taps the transcription process which draws on transcription skills (Kim, 2024). Transcription skills involve spelling and handwriting, which are dissociable constructs and represent a finer grain size (e.g., Kim et al., 2014; Puranik et al., 2008). Third, Word Dictation is a time-limited task, and it is possible that students' working speed may influence their performance. Researchers, particularly those studying computerized assessments, often model accuracy (or item responses) and speed (or response time) as different latent constructs (e.g., Kang et al., 2023; Man et al., 2019). While exploring a sophisticated model that dissects English spelling accuracy and working speed would not be straightforward, especially considering that Word Dictation does not track response time per item, it would be a valuable research endeavor.

Limitations of the Study

The following limitations should be considered when interpreting these findings. First, I only analyzed subsets of Word Dictation items (Items 1-15 in Form A and Items 1-14 in Form B) instead of the entire set of 30 items. This decision was necessary because the majority of students in this study did not reach further items within the time limit. As such, measurement invariance for later items could not be assessed. Furthermore, assessment-level findings should be interpreted only within the context of these item subsets and could differ if all 30 items are included for analysis.

Second, I grouped students with diverse L1s into a single category of "multilingual students," due to the limited number of students with specific languages as their L1. As a result, the findings on DIF and DTF solely apply to the comparison between the multilingual student group, encompassing various L1s, and the English-monolingual student group and should not be extrapolated to students with any particular L1 background.

A third limitation stems from the specific choice of IRT models. In this study, I selected the Rasch model and PCM due to their parsimonious nature. However, these models assume equal discrimination across all items, meaning that items are presumed to differ only in difficulty and not in their ability to distinguish between students of varying ability levels effectively. Under this assumption, regardless of how a student achieves a particular total score, the same score consistently yields the same ability estimate (de Ayala, 2022; Embreston & Reise, 2000). For instance, if one student correctly answers three CVC words in Word Dictation and another student answers three CVCC words, they receive the same ability estimate (based on WSC scores). However, Word Dictation items, with varying complexity levels, might indeed exhibit different discriminatory powers. Consequently, the item response data might have been better explained by other IRT models such as the 2PL model.

Fourth, I examined relatively small sample sizes of students, much smaller than many other measurement invariance studies in education and psychology (Alatli, 2020). As a result, the statistical power of DIF analysis was low, falling below the desired level of .80 (Cohen, 1988). Specifically, post-hoc power to detect true NCDIF values for each item ranged from .054 to .787 (note that due to computational complexity, I could not compute the power of detecting DIF using the IRT likelihood ratio test). Further, NCDIF power was especially low for the items

found to be non-DIF compared to those identified as DIF, suggesting that some items with true DIF could have been incorrectly classified as having no DIF in this study. Thus, the current findings should be interpreted with great caution, and replication with larger samples is needed.

Recommendations for Future Research

Considering these limitations, I propose several directions for future work. First, replicating the current study within a similar population (i.e., linguistically diverse students primarily in Grades 1–3 identified as struggling with English writing) is needed. Replication is crucial to validate the current findings, as the performance of IRT likelihood ratio tests can be influenced by a number of factors including the sample size and the percentage and balance of true DIF items (Wang et al., 2022). Additionally, DFIT statistics are considered suboptimal when the sample size is not large enough (Chalmers, 2018), as evidenced by the inadequate NCDIF power observed in this study.

In such replications, it would be advantageous to incorporate students' item responses for the complete set of 30 items. Then, researchers could assess DIF for later items that were not addressed in this study and also offer clearer insights into measurement invariance at the entire assessment level. To achieve this, researchers may consider having students complete all 30 items with no time limit.

Furthermore, if future researchers choose to use the IRT modeling, they may consider exploring different methods for selecting anchor items. In this study, in the absence of predefined anchor items, I followed the procedure outlined by Tay et al. (2015) to identify DIF-free items and use them as anchor items. However, various other methods for identifying anchor items exist, which are known to result in differing DIF results (Kopf et al., 2015; Wang & Yeh, 2003).

Second, future measurement invariance studies focusing on CLS with a more advanced analysis strategy are needed. Specifically, researchers may reveal specific score levels within an item that contribute to the DIF effect. In this study, the analysis relied on the omnibus measure of DIF for polytomous items, which does not specify the particular score levels associated with the observed DIF (Penfield et al., 2008). That is, although I found DIF in the item "PAT" scored for CLS, I could not determine whether it occurred, for example, between the probability of getting 2 CLS compared to 1 CLS, or between 3 CLS compared to 2 CLS. Similarly, I couldn't identify which specific combination of adjacent letters contributed to the observed DIF. Moving forward, researchers could investigate specific CLS score levels and letter combinations that lead to noticeable language differences in students' item responses, and in this endeavor, using a framework called differential step functioning (Penfield, 2007) would be beneficial.

Moreover, in future measurement invariance investigations focusing on CLS, a different approach to selecting anchor items might be needed. In the current analysis, I initially assessed DIF using WSC scores and then analyzed CLS scores, using the items identified as non-DIF based on WSC scores as anchor items. An implicit assumption underlying this approach to selecting anchor items is that items without DIF in WSC scoring would also not exhibit DIF in CLS. In other words, this approach assumes that an item showing no DIF in WSC would behave similarly in CLS. This assumption was grounded in the idea that CLS, by assigning partial credit for partially correct responses, can more accurately and nuancedly capture students' writing skills, thus making it less prone to DIF compared to WSC scoring. Additionally, this approach aligns with the use of the omnibus DIF measure for polytomous data, as mentioned above, which evaluates DIF by collapsing all score categories within an item (e.g., collapsing the probabilities

of scoring 1 CLS versus 0, 2 CLS versus 1, 3 CLS versus 2, and so on, depending on word length), rather than isolating DIF effect at each specific score level (e.g., DIF in the probability of scoring 2 CLS versus 1 CLS in an item). Based on this, at least within the current study context, I deemed it reasonable to treat non-DIF items from WSC as DIF-free items, and therefore suitable anchor items for the CLS analysis. However, if future researchers are interested in assessing DIF at specific CLS score levels within an item, a more nuanced approach (e.g., initiating the DIF analysis directly with CLS scores and selecting anchor items drawn from those scores) may be beneficial. It remains possible that an item without DIF in WSC could still exhibit DIF at certain CLS score levels, meaning that non-DIF items in WSC scoring might not necessarily be DIF-free in CLS scoring.

Third, this study focused solely on one specific task, Word Dictation, among a variety of English writing CBM tasks available. This choice was intentional because (a) Word Dictation consists of discrete items, making it suitable for item-level analysis, compared to tasks where students generate their own writing, and (b) many teachers in the large RCT chose this measure as a progress monitoring tool, indicating its usefulness for young students struggling with writing. However, future research should explore potential measurement invariance with other types of CBM writing tasks. This broader investigation will contribute to more accurate screening and progress monitoring practices in writing, contributing to better identification and support for all student populations.

Examining sentence and passage-level CBM tasks holds particular importance because many students continue to face writing difficulties in upper elementary grade levels and beyond (National Center for Education Statistics, 2003; 2021). For upper-grade students, passage-level

writing CBM tasks (Campbell, 2010; Ritchey et al., 2016), scored with correct minus incorrect word sequences (Espin et al., 2000; Espin et al., 2004; Romig et al., 2017; Truckenmiller et al., 2020), have traditionally been considered to provide technically rigorous and useful data. Evaluating measurement invariance for such tasks and scoring metrics can be more complex, as determining what qualifies an "item" within sentence- or passage-level tasks may be less straightforward, unlike word-level items, which are more clearly defined. Addressing this complexity would be an important next step.

Fourth, my investigation centered on two particular forms of Word Dictation, but future research should assess the measurement invariance of other available alternate forms. Word Dictation serves as a valuable tool for monitoring students' growth in their overall writing skills and can inform instructional decisions. This way of using Word Dictation reflects the original purpose of developing the CBM framework (Deno, 1985; Fuchs et al., 2010) and has been shown to be effective in improving writing outcomes of students identified with intensive early writing needs (Choi et al., 2023; McMaster et al., 2024). To help students with various language backgrounds benefit from this practice, ensuring the measurement invariance of alternate forms is essential.

Fifth, if larger sample sizes (assessment length and student sample jointly) are available, researchers may consider using the 2PL model to evaluate measurement invariance within the IRT framework. The 2PL model does not assume equal discrimination across items, a condition not met in many assessments (Stemler & Naples, 2021), probably including Word Dictation. Moreover, by adopting the 2PL model, researchers could examine non-uniform DIF (e.g., either multilingual or English-monolingual student group demonstrates a higher probability of getting

the item correct up to a certain level of the latent trait, after which this trend reverses), allowing for a more nuanced understanding of DIF.

Last but most importantly, in future efforts assessing measurement invariance concerning language status, it is essential to move beyond comparing the multilingual student group, encompassing various L1s, with the English monolingual student group. Instead, researchers should focus on specific L1 contexts. The detection and severity of DIF could vary depending on these specific L1 backgrounds; prior research identified that the effect sizes of DIFs varied depending on students' ethnicity (Koo et al., 2014), an aspect related to language. Researchers should also strive to collect data on students' language characteristics and experiences, such as their proficiency in both their L1 and English (beyond the English Learner service eligibility), the language support they receive, and how their L1 is used in their home and community. This emphasis is especially critical when investigating the influence of cross-linguistic transfer, a crucial factor to consider when exploring the underlying reasons for DIF. Without knowing the linguistic characteristics of the student's L1, we cannot speculate on the patterns of cross-linguistic transfer. Similarly, without understanding how much the student has developed grapheme-phoneme knowledge in their L1, we cannot anticipate the likelihood of cross-linguistic transfer occurring.

To provide initial insights into how future investigations exploring the influence of cross-linguistic transfer in explaining DIF within the specific L1 context may or may not reveal, I briefly outline the preliminary analysis of item responses from Spanish-English bilingual students ($N = 33$) in this study. Specifically, I analyzed their actual responses to the item "PAT," which showed a medium-sized DIF. Upon comparing the item responses from English-

monolingual students, I observed some noticeable differences in the types of errors made. Among Spanish-English bilingual students, the most common error type (33%) was replacing ‘a’ with ‘e’ (writing *pet*, which is a homophone – word having the same or similar pronunciation but different meaning or spelling – of *pet*), followed by consonant/vowel replacement by other consonants/vowels (18.5%), writing random or non-readable letters (18.5%), vowel omission (11.1%), providing no response (11.1%), adding word-final ‘e’ (writing *pate*; 3.7%), and reversal of letters ‘p’ and ‘q’ (writing *qat*; 3.7%). Conversely, among English-monolingual students, the most frequent error type was writing random or non-readable letters (23.4%), followed by providing no response (17.2%), consonant/vowel replacement by other consonants/vowels (15.6%), vowel omission (14.8%), replacing ‘a’ with ‘e’ (10.2%), the reversal of ‘p’ and ‘q’ (7.8%), and adding word-final ‘e’ (3.1%).

The errors observed among Spanish-English bilingual students do not align precisely with known types of English spelling errors indicative of Spanish influence (Bahr et al., 2015; Linan-Thompson & Meline, 2022; Rubin & Carlan, 2005; Wolters & Kim, 2024). Nevertheless, it is surprising and may warrant further investigation that, within this study context, phonological errors (e.g., replacement of ‘a’ and ‘e’) were more commonly found among Spanish-English bilingual students, whereas non-predicted errors (neither a phonological nor an orthographic error; e.g., providing random or non-readable letters, skipping the item) were more prevalent in English-monolingual students. No clear conclusions can be drawn, and it is not appropriate to directly link these error patterns to DIF results; again, in the current DIF analysis, the multilingual student group, encompassing different L1s, was compared with the English-

monolingual student group. However, future studies that conduct DIF analysis and subsequent item review, focusing on Spanish-English bilingual students, would provide deeper insights.

Implications for Practice

Challenges in English writing development faced by many students (e.g., National Center for Education Statistics, 2003; 2021), including multilingual students (National Center for Education Statistics, 2012), are of significant concern. To address this concern, educators have implemented diverse and insightful strategies, including adopting research-based writing interventions (Datchuk & Kubina, 2012; Graham et al., 2012; McMaster et al., 2018) and devising innovative writing instruction in their classrooms. Yet, appropriate assessment–a critical element in supporting students' writing development – still requires greater focus. Without appropriate writing assessments, improvements in students' writing skills, as well as our ability as researchers and educators to detect these improvements, may be limited.

In many elementary schools, especially those implementing multi-tiered systems of support, the use of assessments for universal screening and progress monitoring is becoming more prevalent (National Center on Improving Literacy, 2023; Zirkel & Thomas, 2010), with CBM being a frequently used tool. Integrating writing CBM into this system is vital for early identification and providing effective for students' literacy development. Despite the availability of various English writing CBM tasks (e.g., McMaster et al., 2011; Romig et al., 2017), a significant research gap concerning their validity and fairness for multilingual students leads to confusion among educators and stakeholders when using such tools with multilingual students; this issue could also extend to other academic areas like reading and math. Although this study focused on Word Dictation, one type of writing CBM, its findings could have broader practical

implications, offering insights into the valid and fair use of writing assessments to support young, struggling English writers from diverse linguistic backgrounds.

First, the finding that Word Dictation Forms A and B function similarly across multilingual and English-monolingual student groups is encouraging for educators—with the caveat that this finding is based on the first 14 or 15 items. When educators use the measure for screening purposes, such as determining which students might need additional support to meet end-of-year expectations, educators often focus on the total scores. When the focus is on the total scores, educators can choose to use either the WSC or CLS scores, as measurement invariance has been established for both scoring metrics. WSC is straightforward but could often provide limited insights into the students' writing skills; many struggling, young writers might write few words correctly. Then, using CLS could be a better option as it can better distinguish between students with differing levels of writing skills.

Second, when educators intend to use Word Dictation data to guide their instruction (e.g., determining in which content areas students need additional instruction or understanding why a student is struggling with writing), examining item-level results can provide additional insights. Based on findings that DIF detected in WSC scores was alleviated in CLS scores, it is recommended that educators prioritize CLS scores over WSC scores. For example, educators may pay attention to the specific words where students obtained low or high CLS scores. Importantly, in such item-level analyses, I recommend that educators be aware that certain items (e.g., PAT in Form A) could be problematic when assessing the multilingual student's responses; some multilingual students may receive lower CLS scores on these items, which may not accurately reflect their true English transcription skills.

Third, it is essential for educators and stakeholders to understand (at levels appropriate to their roles) the importance of technical adequacy in defining quality assessment (for other essential elements to consider, such as users, decisions, and content, see Truckenmiller et al., 2024, for example). They should understand the types and levels of technical evidence supporting writing assessments and use that knowledge to inform their selection of specific tools. In such consideration, they should consider the evidence regarding measurement invariance related to students' language status. This study confirms the measurement invariance of Word Dictation; however, many other writing CBM tasks have not undergone measurement invariance evaluation. This absence of measurement invariance evaluation does not prevent the use of such tasks for multilingual student populations if other technical quality aspects (e.g., criterion-related validity) are well-supported; this is particularly so because decisions based on CBM are usually low stakes (Michigan Assessment Consortium, 2020). Nonetheless, it is important to exercise caution. When using writing measures that have not undergone prior measurement invariance evaluation, educators and stakeholders should be aware that the assessment results might not offer valid data for some multilingual students.

Last, educators may find it beneficial to delve into students' actual responses to each item, moving beyond merely relying on numeric scores. Students' written products in Word Dictation can serve as a rich source of information that reveals students' strengths and weaknesses in writing, particularly for multilingual students struggling with English writing. For instance, educators may notice distinct patterns in the student's English spelling errors, potentially indicating cross-linguistic transfer. Even without knowledge of the student's first language, educators can use resources like lists of common English spelling errors influenced by

Spanish, as observed among Spanish-English bilingual students (e.g., Bahr et al., 2015; Wolter & Kim, 2024). By recognizing these influences through Word Dictation responses, ideally supplemented with communication with the student's family or guardians to understand the student's linguistic background, educators may be able to tailor instruction effectively.

Conclusion

Effective educational support for struggling beginning writers requires appropriate assessment practices, backed by rigorous evidence suggesting that the assessment tools offer valid and reliable data. Extensive efforts have been made in the research community to provide technical adequacy evidence for various English writing CBMs. However, these efforts have not sufficiently prioritized multilingual students learning to write in English, raising doubts about whether existing English writing CBM tools offer valid information for these students. This dissertation aimed to fill this gap by focusing on Word Dictation, a specific type of writing CBM, designed to assess English transcription skills. The study investigated whether two widely used scores for Word Dictation, WSC and CLS, demonstrate measurement invariance across multilingual and English-monolingual student groups. Within the context of this study, results supported the establishment of measurement invariance at both item and assessment levels. Some items did display differential functioning, but they did so in varying directions rather than systematically favoring or disfavoring any specific group. When comparing WSC and CLS, certain items flagged as DIF in WSC were no longer flagged in CLS, suggesting that CLS may be less prone to DIF. Given the absence of prior assessment of measurement invariance of Word Dictation regarding students' language status, direct comparisons with existing literature are not feasible. Nonetheless, the current findings align with extant knowledge, indicating the potential

and technical rigor of Word Dictation for struggling beginning writers, especially when scored with CLS. Future studies should replicate the current analysis in similar populations, ideally using the entire set of Word Dictation items, to confirm or disconfirm the current study's conclusions. Furthermore, upcoming research must focus on multilingual students with specific native languages rather than treating them as a homogenous group. Despite its limitations, this dissertation may contribute to the valid and fair use of CBMs to support English writing development in young students facing intensive needs and coming from diverse language backgrounds.

References

Adriansen, H. K., Juul-Wiese, T., Møller Madsen, L., Saarinen, T., Spangler, V., & Waters, J. L. (2022). Emplacing English as lingua franca in international higher education: A spatial perspective on linguistic diversity. *Population Space and Place, 29*(2), e2619. https://doi.org/10.1002/psp.2619

Ahmed, Y., & Wagner, R. K. (2020). A "simple" illustration of a joint model of reading and writing using meta-analytic structural equation modeling (MASEM). *Reading-Writing Connections: Towards Integrative Literacy Science,* 55-75. https://doi.org/10.1007/978-3-030-38811-9_4

Alatlı, B. (2020). Cross-cultural measurement invariance of the items in the science literacy test in the programme for international student assessment (PISA-2015). *International Journal of Education and Literacy Studies, 8*(2), 16-27. https://doi.org/10.7575/aiac.ijels.v.8n.2p.16

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

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Lawrence Erlbaum Associates.

Agresti, A. (2002). Categorical data analysis. Hoboken, NJ: Wiley.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397-438. https://doi.org/10.1080/10705510903008204

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16,* 87–96. https://doi.org/10.1177/014662169201600111

Bahr, R. H., Silliman, E. R., Danzak, R. L., & Wilkinson, L. C. (2015). Bilingual spelling patterns in middle school: It is more than transfer. *International Journal of Bilingual Education and Bilingualism, 18*(1), 73-91. https://doi.org/10.1080/13670050.2013.878304

Banks, K. (2015). An introduction to missing data in the context of differential item functioning. *Practical Assessment, Research and Evaluation, 20*(12), 1-10. https://doi.org/10.7275/fpg0-5079

Belzak, W. C. (2020). Testing differential item functioning in small samples. *Multivariate Behavioral Research, 55(*5), 722-747. https://doi.org/10.1080/00273171.2019.1671162

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Berninger, V., & Amtmann, D. (2003). Preventing written expression disabilities through early and continuing assessment and intervention for handwriting and/or spelling problems: Research into practice. In H. L. Swanson, K. Harris, & S. Graham (Eds.), *Handbook of research on learning disabilities* (pp. 345–363). New York, NY: Guilford Press.

Berninger, V. W., & Rutberg, J. (1992). Relationship of finger function to beginning writing: Application to diagnosis of writing disabilities. *Developmental Medicine and Child Neurology, 34*(3), 198–215.

Biancarosa, G., & Snow, C. (2004). *Reading next—A vision for action and research in middle and high school literacy: A report to Carnegie Corporation of New York.* Washington, DC: Alliance for Excellence in Education.

Berninger, V. W. (1999). Coordinating transcription and text generation in working memory during composing: Automatized and constructive processes. *Learning Disability Quarterly, 22*(2), 99–112. https://doi.org/10.2307/1511269

Berninger, V. W., & Winn, W. D. (2006). Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 96 – 114). New York, NY: Guilford Press.

Burns, M. K., Jimerson, S. R., VanDerHeyden, A. M., & Deno, S. L. (2016). Toward a unified response-to-intervention model: Multi-tiered systems of support. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of response to intervention* (pp. 719–732). Springer.

Butvilofsky, S. A., Escamilla, K., Gumina, D., & Silva Diaz, E. (2021). Beyond monolingual reading assessments for emerging bilingual learners: Expanding the understanding of biliteracy assessment through writing. *Reading Research Quarterly, 56*(1), 53–70. https://doi.org/10.1002/rrq.292

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology, 66*(2), 245-276. https://doi.org/10.1111/j.2044-8317.2012.02050.x

Campbell, H. M. (2010). The technical adequacy of curriculum-based measurement passage copying with secondary school English language learners. *Reading & Writing Quarterly, 26*(4), 289–307. https://doi.org/10.1080/10573569.2010.500253

Campbell, H., Espin, C. A., & McMaster, K. (2013). The technical adequacy of curriculum-based writing measures with English learners. *Reading and Writing, 26*(3), 431–452. https://doi.org/10.1007/s11145-012-9375-6

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*(3), 253–260. https://doi.org/10.1177/014662168801200304

Caravolas, M., Hulme, C., & Snowling, M. J. (2001). The foundations of spelling ability: Evidence from a 3-year longitudinal study. *Journal of Memory and Language, 45*(4), 751–774. https://doi.org/10.1006/jmla.2000.2785

Carroll, J. B. (1964). *Language and thought.* Englewood Cliffs, NJ: Prentice-Hall, Inc.

Cervantes, V. H. (2017). DFIT: An R Package for Raju's differential functioning of items and tests framework. *Journal of Statistical Software, 76*(5), 1–24. https://doi.org/10.18637/jss.v076.i05

Choi, S., Shanahan, E., An, J., & McMaster, K. L. (2023). Monitoring elementary students' progress using Word Dictation: Technical features of slope and growth analysis. *Assessment for Effective Intervention, 48*(4), 201-210. https://doi.org/10.1177/15345084231182718

Choi, S., McMaster, K. L., Kohli, N., Shanahan, E., Birinci, S., An, J., Duesenberg-Marshall, M., & Lembke, E. S. (2024). Longitudinal effects of data-based instructional changes for

students with intensive learning needs: A piecewise linear–linear mixed-effects modeling approach. *Journal of Educational Psychology, 116*(4), 608–628. https://doi.org/10.1037/edu0000853

Chung, S. C., Chen, X., & Geva, E. (2019). Deconstructing and reconstructing cross-language transfer in bilingual reading development: An interactive framework. *Journal of Neurolinguistics, 50,* 149-161. https://doi.org/10.1016/j.jneuroling.2018.01.003

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2018). Model-based measures for detecting and quantifying response bias. *Psychometrika, 83,* 696-732. https://doi.org/10.1007/s11336-018-9626-9

Chalmers, R. P. (2023). A unified comparison of IRT-based effect sizes for DIF investigations. *Journal of Educational Measurement, 60*(2), 318-350. https://doi.org/10.1111/jedm.12347

Chung, S. C., Chen, X., & Deacon, S. H. (2018). The relationship between orthographic processing and spelling in Grade 1 French immersion children. *Journal of Research in Reading, 41*(2), 290-311. http://dx.doi.org/10.1111/1467-9817.12104

Clemens, N. H., Keller-Margulis, M. A., Scholten, T., & Yoon, M. (2016). Screening assessment within a multi-tiered system of support: Current practices, advances, and next steps. In S. R. Jimerson, M. K., Burns, & A. M. VanDerHeyden (Eds.), *Handbook of response to intervention* (pp. 187–213). Springer. https://doi.org/10.1007/978-1-4899-7568-3_12

Coady, J. (1997). L2 vocabulary acquisition: A synthesis of research. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 225-237). New York: Cambridge University Press.

Cohen, J. (1988). *Statistical power analyses for the behavioral sciences* (2nd edition). Hillsdale, NJ: Erlbaum.

Connor, U. (1996). *Contrastive rhetoric: Cross-cultural aspects of second-language writing*. New York: Cambridge University Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. https://doi.org/10.1007/BF02310555

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49*(2), 222–251. https://doi.org/10.3102/00346543049002222

Datchuk, S. M., & Kubina, R. M. (2012). A review of teaching sentence-level writing skills to students with writing difficulties and learning disabilities. Remedial and Special Education, 34, 180–192. https://doi.org/10.1177/0741932512448254

Diaz, R. M., & Klingler, C. (1991). Towards an explanatory model of the interaction between bilingualism and cognitive development. In E. Bialystok (Ed.), *Language processing in bilingual children* (pp. 167–192). Cambridge, UK: Cambridge University Press.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95*(1), 134–135. https://doi.org/10.1037/0033-2909.95.1.134

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: a procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*(3), 363-373. https://doi.org/10.1037/0021-9010.68.3.363

de Ayala, R. J. (2022). *The theory and practice of item response theory.* The Guilford Press.

DeMars, C. E. (2011). An analytic comparison of effect sizes for differential item functioning. *Applied Measurement in Education, 24*(3), 189-209. https://doi.org/10.1080/08957347.2011.580255

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219-232. https://doi.org/10.1177/001440298505200303

Deno, S. L., Mirkin, P. K., & Marston, D. (1980). *Relationships among simple measures of written expression and performance on standardized achievement tests* (Report No. 22). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods, 23*(1), 138–149. https://doi.org/10.1037/met0000121

Ehri, L. (2000). Learning to read and learning to spell: Two sides of a coin. *Topics in Language Disorders, 20,* 19–36. https://doi.org/10.1097/00011363-200020030-00005

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum Associates, Inc., Publishers.

Espin, C. A., & Deno, S. (2016). Conclusion: Oral reading fluency or reading aloud from text: An analysis through a unified view of construct validity. In K. D. Cummings & Y. Petscher

(Eds.), *The fluency construct: Curriculum-based measurement concepts and applications* (pp. 365–384). Springer.Espin, C. A., Weissenburger, J. W., & Benson, B. J. (2004). Assessing the writing performance of students in special education. Exceptionality, 12(1), 55– 66. https://doi.org/10.1207/s15327035ex1201_5

Espin, C., Shin, J., Deno, S. L., Skare, S., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *Journal of Special Education, 34*(3), 140 –153. https://doi.org/10.1177/002246690003400303

Espin, C., Wallace, T., Campbell, H., Lembke, E. S., Long, J. D., & Ticha, R. (2008). Curriculum-based measurement in writing: Predicting the success of high-school students on state standards tests. *Exceptional Children, 74*(2), 174–193. https://doi.org/10.1177/00144029080740020

Farrington, A. L., Lonigan, C. J., Phillips, B. M., Farver, J. M., & McDowell, K. D. (2015). Evaluation of the utility of the revised Get Ready to Read! for Spanish-speaking English-language learners through differential item functioning analysis. *Assessment for Effective Intervention, 40*(4), 216–227. https://doi.org/10.1177/1534508415577468

Fitzgerald, J., & Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist, 35*(1), 39-50. https://doi.org/10.1207/S15326985EP3501_5

Fitzgerald, J., Olson, C. B., García, S. G., & Scarcella, R. (2014). Assessing bilingual students' writing. In A. Clinton (Ed.), *Integrated assessment of the bilingual child* (pp. 215–240). American Psychological Association.

Flores, N., & Rosa, J. (2015). Undoing appropriateness: Raciolinguistic ideologies and language diversity in education. *Harvard Educational Review, 85*(2), 149-171. https://doi.org/10.17763/0017-8055.85.2.149

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education, 41*, 121–139. https://doi.org/10.1177/00224669070410020101

Frisby, C. (2016). An empirical comparison of the words spelled correctly and correct letter sequence spelling scoring methods in third-and fourth-grade classrooms. *Journal of Applied School Psychology, 32*(2), 101–121. https://doi.org/10.1080/15377903.2016.1151847

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*, 188–192. https://doi.org/10.1080/02796015.2004.12086241

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57*(6), 488–500. https://doi.org/10.1177/001440299105700603

Fuchs, L. S., & Vaughn, S. (2012). Responsiveness-to-Intervention: A decade later. *Journal of Learning Disabilities, 45*(3), 195–203. https://doi.org/10.1177/0022219412442150

Fuchs, D., Fuchs, L. S., & Stecker, P. M. (2010). The "blurring" of special education in a new continuum of general education placements and services. *Exceptional Children, 76*(3), 301-323. https://doi.org/10.1177/001440291007600304

Furnes, B., & Samuelsson, S. (2009). Preschool cognitive and language skills predicting kindergarten and grade 1 reading and spelling: A cross-linguistic comparison. *Journal of Research in Reading, 32*(3), 275–292. https://doi.org/10.1111/j.1467-9817.2009.01393.x

Genesee, F., Geva, E., Dressler, C. & Kamil, M. L. (2006). Cross-linguistic relationships in second-language learners. In D. August & T. Shanahan (Eds.). *Developing reading and writing in second-language learners: Lessons from the report of the National Literacy Panel on language-minority children and youth* (pp. 61-93). New York, NY: Routledge.

Geva, E., & Ryan, E. B. (1993). Linguistic and cognitive correlates of academic skills in first and second language. *Language Learning*, *43*(1), 5–42. https://doi.org/10.1111/j.1467-1770.1993.tb00171.x

Goodrich, J. M., Lonigan, C. J., & Alfonso, S. V. (2019). Measurement of early literacy skills among monolingual English-speaking and Spanish-speaking language-minority children: A differential item functioning analysis. *Early Childhood Research Quarterly*, *47*, 99–110. https://doi.org/10.1016/j.ecresq.2018.10.007

Graham, K. M., & Eslami, Z. R. (2020). Does the simple view of writing explain L2 writing development?: A meta-analysis. *Reading Psychology*, *41*(5), 485-511. https://doi.org/10.1080/02702711.2020.1768989

Graham, K. M., & Eslami, Z. (2022). Using the simple view of writing for explaining English L2 writing variation. *Reading Psychology, 43*(7), 523-540. https://doi.org/10.1080/02702711.2022.2126573

Graham, S., Collins, A. A., & Rigby-Wills, H. (2017). Writing characteristics of students with learning disabilities and typically achieving peers: A meta-analysis. *Exceptional Children, 83*(2), 199-218. https://doi.org/10.1177/0014402916664070

Graham, S., Fishman, E. J., Reid, R., & Hebert, M. (2016). Writing characteristics of students with attention deficit hyperactive disorder: A meta-analysis. *Learning Disabilities Research & Practice, 31*(2), 75-89. https://doi.org/10.1111/ldrp.12099

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology, 104*(4), 879–896. https://doi.org/10.1037/a0029185

Graham, S., Hebert, M., Fishman, E., Ray, A. B., & Rouse, A. G. (2020). Do children classified with specific language impairment have a learning disability in writing? A meta-analysis. *Journal of Learning Disabilities, 53*(4), 292-310. https://doi.org/10.1177/0022219420917338

Graham, S., Kim, Y. S., Cao, Y., Lee, J., Tate, T., Collins, P., Choi, M., Chung, H. Q., & Olson, C. B. (2023). A meta-analysis of writing treatments for students in grades 6–12. *Journal of Educational Psychology*, *115*(7), 1004-1927. https://doi.org/10.1037/edu0000819

Grigorenko, E. L., Compton, D. L., Fuchs, L. S., Wagner, R. K., Willcutt, E. G., & Fletcher, J. M. (2020). Understanding, educating, and supporting children with specific learning disabilities: 50 years of science and practice. *American Psychologist, 75*(1), 37–51. https://doi.org/10.1037/amp0000452

Grosjean, F. (2010). *Bilingual: Life and reality*. Cambridge, MA: Harvard University Press.

Hampton, D. D., & Lembke, E. S. (2016). Examining the technical adequacy of progress monitoring using early writing curriculum-based measures. *Reading & Writing Quarterly, 32*(4), 336–352. https://doi.org/10.1080/10573569.2014.973984

Harrison, E. (2020). *Package "finalfit"*. https://cran.rproject.org/web/packages/finalfit/finalfit.pdf

Helman, L., Ittner, A. C., & McMaster, K. L. (2020). *Assessing language and literacy with bilingual students.* The Guilford Press.

Holland, W. P., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. W., & Wainer, H. (2012). *Differential item functioning.* Routledge.

Hooper, S. R., Costa, L., McBee, M., Anderson, K. L., Yerby, D. C., Knuth, S. B., et al. (2011). Concurrent and longitudinal neuropsychological contributors to written language expression in first and second grade students. *Reading and Writing, 24*, 221–252. https://doi.org/10.1007/s11145-010-9263-x

Hopewell, S., & Butvilofsky, S. (2016). Privileging bilingualism: Using biliterate writing outcomes to understand emerging bilingual learners' literacy achievement. *Bilingual Research Journal, 39*(3-4), 324-338. https://doi.org/10.1080/15235882.2016.1232668

Hosp, M. K., Hosp, J. L., & Howell, K. W. (2016). *The ABCs of CBM.* New York, NY: The Guilford Press.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461–473. https://doi.org/10.1093/applin/amp048

Houts, C. R., & Edwards, M. C. (2013). The performance of local dependence measures with psychological data. *Applied Psychological Measurement, 37*(7), 541–562. https://doi.org/10.1177/0146621613491456

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. https://doi.org/10.1037/1082-989X.3.4.424

Hulme, C., Hatcher, P. J., Nation, K., Brown, A., Adams, J., & Stuart, G. (2002). Phoneme awareness is a better predictor of early reading skill than onset-rime awareness. *Journal of Experimental Child Psychology*, *82*(1), 2-28. https://doi.org/10.1006/jecp.2002.2670

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. *School Psychology Review, 36*, 582–600. https://doi.org/10.1080/02796015.2007.12087919

Jewell, J., & Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *School Psychology Review*, *34*(1), 27–44. https://doi.org/10.1080/02796015.2005.12086273

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika, 75*, 393-419.

Jonson, J. L., & Geisinger, K. (2022). *Fairness in educational and psychological testing: Examining theoretical, research, practice, and policy implications of the 2014 standards.* American Educational Research Association.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications.* Chicago, IL: SPSS Inc.

Juel, C., Griffith, P. L., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology, 78*(4), 243–255. https://doi.org/10.1037/0022-0663.78.4.243

Jung, P. G., McMaster, K. L., Kunkel, A. K., Shin, J., & Stecker, P. M. (2018). Effects of data–based individualization for students with intensive learning needs: A meta–analysis. *Learning Disabilities Research & Practice, 33*(3), 144-155. https://doi.org/10.1111/ldrp.12172

Kang, I., Jeon, M., & Partchev, I. (2023). A latent space diffusion item response theory model to explore conditional dependence between responses and response times. *Psychometrika, 88*(3), 830-864. https://doi.org/10.1007/s11336-023-09920-x

Kaufman, A. S., & Kaufman, N. L. (2014). *Kaufman test of education achievement* (3rd ed.). Bloomington: NCS Pearson.

Keller-Margulis, M. A., Mercer, S. H., & Thomas, E. L. (2016). Generalizability theory reliability of written expression curriculum-based measurement in universal screening. *School Psychology Quarterly, 31(*3), 383–392. https://doi.org/10.1037/spq0000126

Keller-Margulis, M., Payan, A., Jaspers, K. E., & Brewton, C. (2016). Validity and diagnostic accuracy of written expression curriculum-based measurement for students with diverse

language backgrounds. *Reading & Writing Quarterly*, *32*(2), 174-198. https://doi.org/10.1080/10573569.2014.964352

Kent, S. C., & Wanzek, J. (2016). The relationship between component skills and writing quality and production across developmental levels: A meta-analysis of the last 25 years. *Review of Educational Research, 86*(2), 570-601. https://doi.org/10.3102/0034654315619491

Kent, S. C., Wanzek, J., & Al Otaiba, S. (2017). Reading instruction for fourth-grade struggling readers and the relation to student outcomes. *Reading & Writing Quarterly*, *33*(5), 395-411. https://doi.org/10.1080/10573569.2016.1216342

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of School Psychology, 52*(4), 377-405. https://doi.org/10.1016/j.jsp.2014.06.002

Kim, D., De Ayala, R. J., Ferdous, A. A., & Nering, M. L. (2011). The comparative performance of conditional independence indices. *Applied Psychological Measurement, 35*(6), 447-471. https://doi.org/10.1177/0146621611407909

Kim, J., & Oshima, T. C. (2013). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement, 73*(3), 458-470. https://doi.org/10.1177/0013164412467033

Kim, Y.-S. G. (2011). Considering linguistic and orthographic features in early literacy acquisition: Evidence from Korean. *Contemporary Educational Psychology*, *36*(3), 177–189. https://doi.org/10.1016/j.cedpsych.2010.06.003

Kim, Y.-S. G. (2020). Interactive dynamic literacy model: An integrative theoretical framework for reading and writing relations. In R. Alves, T. Limpo, & M. Joshi (Eds.), *Reading-writing connections: Towards integrative literacy science* (pp. 11–34). Springer. https://doi.org/10.1007/978-3-03038811-9_2

Kim, Y.-S. G. (2024). Writing fluency: Its relations with language, cognitive, and transcription skills, and writing quality using longitudinal data from kindergarten to grade 2. *Journal of Educational Psychology, 116*(4), 590–607. https://doi.org/10.1037/edu0000841

Kim, Y.-S. G., & Graham, S. (2022). Expanding the Direct and Indirect Effects Model of Writing (DIEW): Reading–writing relations, and dynamic relations as a function of measurement/dimensions of written composition. *Journal of Educational Psychology, 114*(2), 215–238. https://doi.org/10.1037/edu0000564

Kim, Y.-S. G., & Petscher, Y. (2023). Do spelling and vocabulary improve classification accuracy of children's reading difficulties over and above word reading?. *Reading Research Quarterly, 58*(2), 240-253. https://doi.org/10.1002/rrq.496

Kim, Y.-S. G., & Schatschneider, C. (2017). Expanding the developmental models of writing: A direct and indirect effects model of developmental writing (DIEW). *Journal of Educational Psychology, 109(*1), 35–50. https://doi.org/10.1037/edu0000129

Kim, Y.-S. G., Al Otaiba, S., Wanzek, J., & Gatlin, B. (2015). Toward an understanding of dimensions, predictors, and the gender gap in written composition. *Journal of Educational Psychology, 107*(1), 79–95. https://doi.org/10.1037/a0037210

Kim, Y.-S. G., Gatlin, B., Al Otaiba, S., & Wanzek, J. (2018). Theorization and an empirical investigation of the component-based and developmental text writing fluency construct.

Journal of Learning Disabilities, 51(4), 320–335. https://doi.org/10.1177/0022219417712016

Kim, Y.-S. G., Al Otaiba, S., Sidler, J. F., Greulich, L., & Puranik, C. (2014). Evaluating the dimensionality of first-grade written composition. *Journal of Speech, Language, and Hearing Research, 57,* 199 –211.

Kim, Y.-S. G., Al Otaiba, S., Puranik, C., Folsom, J. S., Greulich, L., & Wagner, R. K. (2011). Componential skills of beginning writing: An exploratory study. Learning and Individual Differences, 21(5), 517–525. https://doi.org/10.1016/j.lindif.2011.06.004

Koda, K. (2008). Impacts of prior literacy experience on second-language learning to read. In K. Koda, & A. M. Zehler (Eds.). *Learning to read across languages: Crosslinguistic relationships in first- and second-language literacy development* (pp. 68–96). Mahwah, NJ: Routledge.

Koo, J., Becker, B. J., & Kim, Y. S. (2014). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. *Language Testing, 31(*1), 89-109. https://doi.org/10.1177/0265532213496097

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement, 75*(1), 22–56. https://doi.org/10.1177/0013164414529792

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers*. Ann Arbor: University of Michigan Press.

Lanauze, M., & Snow, C. (1989). The relation between first-and second-language writing skills: Evidence from Puerto Rican elementary school children in bilingual programs.

*Linguistics and Education, 1(*4), 323-339. https://doi.org/10.1016/S0898-5898(89)80005-1

Landis, B. C. (2019). *Formative language assessment for English learners: An exploration of vocabulary diversity indices in writing CBM* [Doctoral dissertation, University of Oregon]. ProQuest Dissertations & Theses Global.

Lembke, E., Deno, S. L., & Hall, K. (2003). Identifying an indicator of growth in early writing proficiency for elementary school students. *Assessment for Effective Intervention, 28*(3–4), 23–35. https://doi.org/10.1177/073724770302800304

Lembke, E. S., McMaster, K. L., Smith, R. A., Allen, A., Brandes, D., & Wagner, K. (2018). Professional development for data-based instruction in early writing: Tools, learning, and collaborative support. *Teacher Education and Special Education, 41*(2), 106-120. https://doi.org/10.1177/0888406417730112

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40(*3), 387-417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Lesaux N. K., Geva, E., Koda, K., Siegel, L. S., & Shanahan, T. (2006). Development of literacy in second-language learners. In D. August & T. Shanahan (Eds.). *Developing reading and writing in second-language learners: Lessons from the report of the National Literacy Panel on language-minority children and youth* (pp. 27-61). New York, NY: Routledge.

Linan-Thompson, S., & Meline, M. (2022). Keys to understanding the writing development of emergent bilingual students. In G. Colón, & T. O. Alsace (Eds.). *Bilingual special education for the 21st century: A new interface* (pp. 179-204). IGI Global.

Lonigan, C. J., Schatschneider, C., Westberg, L., & The National Early Literacy Panel (2008). Identification of children's skills and abilities linked to later outcomes in reading, writing, and spelling. In National Early Literacy Panel (Ed.). *Developing early literacy: A scientific synthesis of early literacy development and implications for intervention* (pp. 55–105). Jessups, ML: National Institute for Literacy & The Partnership for Reading.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (1st ed.). Routledge.

MacWhinney, B. (2005). A unified model of language acquisition. In J. F. Kroll & A. M. B. De Groot (eds.), *Handbook of bilingualism: Psycholinguistic approaches (*pp. 49–67). New York, NY: Oxford University Press.

Magis, D., Beland, S., & Raiche, G. (2020). *Package 'difR'*. https://cran.rproject.org/web/packages/difR/difR.pdf

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1), 85–104. https://doi.org/10.1191/0265532202lt221oa

Man, K., Harring, J. R., Jiao, H., & Zhan, P. (2019). Joint modeling of compensatory multidimensional item responses and response times. *Applied Psychological Measurement, 43*(8), 639-654. https://doi.org/10.1177/0146621618824853

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139–160. https://doi.org/10.1111/j.1745-3984.1977.tb00033.x

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), 149-174. https://doi.org/10.1007/BF02296272

Maydeu-Olivares, A., & Garcia-Forero, C. (2010). Goodness-of-fit testing. *International Encyclopedia of Education, 7*, 190-196. https://doi.org/10.1016/B978-0-08-044894-7.01333-6

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-Information estimation and goodness-of-fit testing in 2p contingency tables: A unified framework. *Journal of the American Statistical Association, 100*, 1009-1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71,* 713-732.

McBride-Chang, C., Cho, J.-R., Liu, H., Wagner, R. K., Shu, H., Zhou, A., et al. (2005). Changing models across cultures: Associations of phonological and morphological awareness to reading in Beijing, Hong Kong, Korea, and America. *Journal of Experimental Child Psychology, 92*(2), 140–160. https://doi.org/10.1016/j.jecp.2005.03.009

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review, 8,* 299 –325. http://dx.doi.org/10.1007/BF01464076

McCutchen, D. (2006). Cognitive factors in the development of children's writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 115–130). New York, NY: The Guilford Press.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*(1), 100-117.

McMaster, K. L., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. *The Journal of Special Education, 41*, 68–84. https://doi.org/10.1177/00224669070410020301

McMaster, K. L., & Lembke, E. S. (2018). *Data-based instruction in beginning writing: A manual.* Minneapolis, MN: University of Minnesota.

McMaster, K. L., Ritchey, K. D., & Lembke, E. (2011a). Curriculum-based measurement for beginning writers: Recent developments and future directions. In T. E. Scruggs & M. A. Mastropieri (Eds.), *Assessment and intervention: Advances in learning and behavioral disabilities* (Vol. 24). Emerald.

McMaster, K. L., Kunkel, A., Shin, J., Jung, P. G., & Lembke, E. (2018). Early writing intervention: A best evidence synthesis. *Journal of Learning Disabilities, 51*(4), 363-380. https://doi.org/10.1177/0022219417708169

McMaster, K. L., Du, X., Yeo, S., Deno, S. L., Parker, D., & Ellis, T. (2011b). Curriculum-based measures of beginning writing: Technical features of the slope. *Exceptional Children, 77*(2), 185–206. https://doi.org/10.1177/001440291107700203

McMaster, K. L., Lembke, E. S., Shin, J., Poch, A. L., Smith, R. A., Jung, P.-G., Allen, A. A., & Wagner, K. (2020). Supporting teachers' use of data-based instruction to improve students' early writing skills. *Journal of Educational Psychology, 112*(1), 1–21. https://doi.org/10.1037/edu0000358

McMaster, K. L., Lembke, E. S., Shanahan, E., Choi, S., An, J., Schatschneider, C., Duesenberg-Marshall, M., Birinci, S., McCollom, E., Garman, C., & Moore, K. (2024). *Supporting*

teachers' data-based individualization of early writing instruction: An efficacy trial [Manuscript submitted for publication].

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*(4), 728.https://doi.org/10.1037/a0018966

Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading, 34*(1), 114–135. https://doi.org/10.1111/j.1467-9817.2010.01477.x

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11. https://doi.org/10.3102/0013189X018002005\

Michigan Assessment Consortium. (2020). *Early literacy assessment systems that support learning*. https://www.michiganas sessmentconsortium.org/ELAS/

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*(4), 297-334. https://doi.org/10.1177/014662169301700401

Muthen, B. O. (1985). A method for studying the homogeneity of test items with respect to other ´ relevant variables. *Journal of Educational Statistics*, *10*(2), 121–132. https://doi.org/10.3102/10769986010002121

Muthen, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT ´ to external variables. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 213–238). Hillsdale, NJ: Lawrence Erlbaum.

National Center for Education Statistics. (2003). *The Nation's Report Card: Writing 2002* (NCES 2003–529). Institute of Education Sciences, U.S. Department of Education.

National Center for Education Statistics. (2012). *The Nation's Report Card: Writing 2011* (NCES 2012–470). Institute of Education Sciences, U.S. Department of Education.

National Center for Educational Statistics. (2021). *How did U.S. students perform on the most recent assessment?* Retrieved from https://www.nationsreportcard.gov/.

National Center for Education Statistics. (2023). *The condition of education: English Learners in public schools.* Washington, D.C: Institute of Education Sciences, U.S. Department of Education.

National Center on Improving Literacy. (2023). *State of Dyslexia, Screening Policies*. https://improvingliteracy.org/state-of-dyslexia

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common core state standards [English language arts]*.

O'Connor, B. P. (2023). *Package 'EFA.dimensions'.* https://cran.r-project.org/web/packages/EFA.dimensions/EFA.dimensions.pdf

Olinghouse, N. G., & Leaird, J. T. (2009). The relationship between measures of vocabulary and narrative writing quality in second- and fourth-grade students. *Reading and Writing, 22*(5), 545–565. https://doi.org/10.1007/s11145-008-9124-z

Oshima, T. C., & Morris, S. B. (2008). Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice*, *27*(3), 43-50. https://doi.org/10.1111/j.1745-3992.2008.00127.x

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (Vol. 161). Thousand Oaks, CA: Sage.

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel–Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement, 71*(6), 1023-1046. https://doi.org/10.1177/0013164411400734

Pendergast, L. L., von der Embse, N., Kilgus, S. P., & Eklund, K. R. (2017). Measurement equivalence: A non-technical primer on categorical multi-group confirmatory factor analysis in school psychology. *Journal of School Psychology, 60*, 65-82. https://doi.org/10.1016/j.jsp.2016.11.002

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement, 4*4(3), 187–210. https://doi.org/10.1111/j.1745-3984.2007.00034.x

Penfield, R. D., & Lam, T. C. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*(3), 5-15. https://doi.org/10.1111/j.1745-3992.2000.tb00033.x

Penfield, R. D., Alvarez, K., & Lee, O. (2008). Using a taxonomy of differential step functioning to improve the interpretation of DIF in polytomous items: An illustration. *Applied Measurement in Education, 22*(1), 61-78. https://doi.org/10.1080/08957340802558367

Pennington, R. C., & Delano, M. E. (2012). Writing instruction for students with autism spectrum disorders: A review of literature. *Focus on Autism and Other Developmental Disabilities, 27(*3), 158-167. https://doi.org/10.1177/1088357612451318

Poch, A. L., Allen, A. A., & Lembke, E. S. (2019). Scoring measures of word dictation curriculum-based measurement in writing: Effects of incremental administration. *Psychology in the Schools, 56*(5), 702–723. https://doi.org/10.1002/pits.22220

Puranik, C. S., Lonigan, C. J., & Kim, Y.-S. (2011). Contributions of emergent literacy skills to name writing, letter writing, and spelling in preschool children. *Early Childhood Research Quarterly, 26*(4), 465–474. https://doi.org/10.1016/j.ecresq.2011.03.002

Puranik, C. S., Lombardino, L., & Altmann, L. (2008). Assessing the microstructure of written language using a retelling paradigm. *American Journal of Speech Language Pathology, 17,* 107–120. https://doi.org/10.1044/1058-0360(2008/012)

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.1.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495-502. https://doi.org/10.1007/BF02294403

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207. https://doi.org/10.1177/01466216900140020

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: a comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517–529. https://doi.org/10.1037/0021-9010.87.3.517

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*(4), 353–368. https://doi.org/10.1177/014662169501900405

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207-230. https://doi.org/10.3102/10769986004003207

Reno, E. A., & McMaster, K. L. (2024). Measuring linguistic growth in sentence-level writing curriculum-based measures: Exploring complementary scoring methods. *Language, Speech, and Hearing Services in Schools, 55*, 1-16. https://doi.org/10.1044/2023_LSHSS-23-00056

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47(*6), 427–469. https://doi.org/10.1016/j.jsp.2009.07.001

Riazi, M., Shi, L., & Haggerty, J. (2018). Analysis of the empirical research in the Journal of Second Language Writing at its 25th year (1992–2016). *Journal of Second Language Writing, 41*, 41-54. https://doi.org/10.1016/j.jslw.2018.07.002

Ritchey, K. D., McMaster, K. L., Al Otaiba, S., Puranik, C. S., Kim, Y.-S. G., Parker, D. C., & Ortiz, M. (2016). Indicators of fluent writing in beginning writers. In. K. D. Cummings, & Y. Petscher (Eds.), *The fluency construct* (pp. 21–66). New York, NY: Springer.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analysis. *Journal of Statistical Software*, *17*, 1–25.

Rodríguez-Casallas, J. D., Luo, W., & Geng, L. (2020). Measuring environmental concern through international surveys: A study of cross-cultural equivalence with item response theory and confirmatory factor analysis. *Journal of Environmental Psychology, 71,* 101494. https://doi.org/10.1016/j.jenvp.2020.101494

Romig J. E., & Olsen A. A. (2021). Technical features of slopes for curriculum-based measures of secondary writing. *Reading & Writing Quarterly*, 37(6), 535–551. https://doi.org/10.1080/10573569.2020.1860841

Romig, J. E., Therrien, W. J., & Lloyd, J. W. (2017). Meta-analysis of criterion validity for curriculum-based measurement in written language. *The Journal of Special Education, 51*(2), 72-82. https://doi.org/10.1177/0022466916670637

Romig, J. E., Miller, A. A., Therrien, W. J., & Lloyd, J. W. (2021). Meta-analysis of prompt and duration for curriculum-based measurement of written language. *Exceptionality, 29*(2), 133-149. https://doi.org/10.1080/09362835.2020.1743706

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

Rubin, R., & Carlan, V. G. (2005). Using writing to understand bilingual children's literacy development. *The Reading Teacher, 58*(8), 728-739. https://doi.org/10.1598/RT.58.8.3

Salvia, J., Ysseldyke, J., & Witmer, S. (2017). *Assessment in special and inclusive education* (13th ed.). Boston: Cengage Learning.

Sanchez, G. H. (1995). *An investigation of the reliability and validity of bilingual curriculum-based measures* [Doctoral dissertation, California State University]. ProQuest Dissertations & Theses Global.

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

Schnoor, B., & Usanova, I. (2023). Multilingual writing development: Relationships between writing proficiencies in German, heritage language and English. *Reading and Writing, 36*(3), 599-623. https://doi.org/10.1007/s11145-022-10276-4

Scott, C. (2009). Language-based assessment of written expression. In G. A. Troia (Eds.), *Instruction and assessment for struggling writers: Evidence-based practices*. Guilford Press.

Shin, J., & McMaster, K. (2019). Relations between CBM (oral reading and maze) and reading comprehension on state achievement tests: A meta-analysis. *Journal of School Psychology, 73*, 131-149. https://doi.org/10.1016/j.jsp.2019.03.005

Smith, R. A., & Lembke, E. S. (2021). Aspects of technical adequacy of an early-writing measure for English language learners in Grades 1 to 3. *Assessment for Effective Intervention, 47*(1), 59–63. https://doi.org/10.1177/1534508420947157

Smith, R. A., & Lembke, E. S. (2022). Technical adequacy of a spelling curriculum-based measure for English language learners in the first through third grades. *Learning Disability Quarterly, 45(*2), 144–155. https://doi.org/10.1177/0731948720930423

Smith, R. A., Allen, A. A., & Alley, J. (2023). A literature synthesis of curriculum-based measurement in writing for English learners. *Psychology in the Schools*. Advance online publication https://doi.org/10.1002/pits.23121

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum-based measurement to improve student achievement: Review of research. *Psychology in the Schools, 42*(8), 795–819. https://doi.org/10.1002/pits.20113

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*(2), 173-180.

Stemler, S. E., & Naples, A. (2021). Rasch measurement v. Item response theory: knowing when to cross the line. *Practical Assessment, Research & Evaluation, 26*, 11. https://doi.org/10.7275/v2gd-4441

Swaminathan, H., & Rogers, J. H. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods, 18*(1), 3–46. https://doi.org/10.1177/1094428114553062

Tay, L., Newman, D. A., & Vermunt, J. K. (2011). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement

equivalence. *Organizational Research Methods*, *14*(1), 147–176. https://doi.org/10.1177/1094428110366037

Tay, L., Vermunt, J. K., & Wang, C. (2013). Assessing the item response theory with covariate (IRT-C) procedure for ascertaining differential item functioning. *International Journal of Testing, 13*(3), 201-222. https://doi.org/10.1080/15305058.2012.692415

Taylor, R. L. (2003). *Assessment of exceptional students: Educational and psychological procedures* (6th ed.). Boston: Allyn and Bacon.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Thordardottir, E. (2011). The relationship between bilingual exposure and vocabulary development. *International Journal of Bilingualism*, *15*(4), 426-445. https://doi.org/10.1177/1367006911403202

Truckenmiller, A. J., Cho, E., & Troia, G. A. (2022). Expanding assessment to instructionally relevant writing components in middle school. *Journal of School Psychology*, *94,* 28–48. https://doi.org/10.1016/j.jsp.2022.07.002

Truckenmiller, A. J., Cho, E., Bourgeois, S., & Friedman, E. (2024). Uses and misuses of commercial reading assessment: An applied framework for decision making in Grades K through 6. *The Reading Teacher.* https://doi.org/10.1002/trtr.2274

Truckenmiller, A. J., McKindles, J. V., Petscher, Y., Eckert, T. L., & Tock, J. (2020). Expanding curriculum-based measurement in written expression for middle school. *The Journal of Special Education, 54*(3), 133-145. https://doi.org/10.1177/0022466919887150

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1-10. https://doi.org/10.1007/BF02291170

Usanova, I., & Schnoor, B. (2021). Exploring multiliteracies in multilingual students: Profiles of multilingual writing skills. *Bilingual Research Journal, 44*(1), 56–73. https://doi.org/10.1080/15235882.2021.1890649

Valdés, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias*. Westport, CT: Ablex Publishing.

van de Vijver, F. J. R., & Leung, K. (2010). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. Matsumoto, & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 17–45). New York, NY: Cambridge University Press.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3(*1), 4-70.https://doi.org/10.1177/109442810031

Verhoeven, L. T. (1994). Transfer in bilingual development: The linguistic interdependence hypothesis revisited. *Language learning, 44*(3), 381-415. https://doi.org/10.1111/j.1467-1770.1994.tb01112.x

Videen, J., Deno, S., & Marston, D. (1982). *Correct word sequences: A valid indicator of proficiency in written expression* (Report No. 84). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Wagner, K., Smith, A., Allen, A., McMaster, K., Poch, A., & Lembke, E. (2019). Exploration of new complexity metrics for curriculum-based measures of writing. *Assessment for Effective Intervention, 44(*4), 256-266. https://doi.org/10.1177/1534508418773448

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*(6), 479–498. https://doi.org/10.1177/0146621603259902

Wang, W., Liu, Y., & Liu, H. (2022). Testing differential item functioning without predefined anchor items using robust regression. *Journal of Educational and Behavioral Statistics, 47*(6), 666-692. https://doi.org/10.3102/10769986221109208

Wayman, M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41*(2), 85–120. https://doi.org/10.1177/00224669070410020401

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE).

Williams, C., & Lowrance-Faulhaber, E. (2018). Writing in young bilingual children: Review of research. *Journal of Second Language Writing, 42,* 58-69. https://doi.org/10.1016/j.jslw.2018.10.012

Winkes, J., & Schaller, P. (2022). Generalizability of written expression curriculum-based-measurement in the German language: What are the major sources of variability? *Frontiers in Education, 7,* 919756. https://doi.org/10.3389/feduc.2022.919756

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. Honolulu, HI: University of Hawaii Press.

Wolters, A. P., & Kim, Y. S. G. (2024). Crosslinguistic influence on spelling in written compositions: Evidence from English-Spanish dual language learners in primary grades. *Reading and Writing, 37*(5), 1059-1078. https://doi.org/10.1007/s11145-023-10416-4

Woolpert, D. (2016). Doing more with less: the impact of lexicon on dual-language learners' writing. *Reading and Writing, 29*(9), 1865-1887. https://doi.org/10.1007/s11145-016-9656-6

Ye, Y., McBride, C., Yin, L., Cheang, L. M. L., & Tse, C. Y. (2022). A model of Chinese spelling development in Hong Kong kindergarteners. *Journal of Learning Disabilities, 55*(2), 154-167. https://doi.org/10.1177/0022219420979959

Yeo, S. (2010). Predicting performance on state achievement tests using curriculum-based measurement in reading: A multilevel meta-analysis. *Remedial and Special Education, 31*(6), 412–422. https://doi.org/10.1177/0741932508327463

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30(*3), 187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Yen, W. M. (1984). Effect of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145. http://dx.doi.org/10.1177/ 014662168400800201

Zieky, M. (2016). Fairness in test design and development. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement (*pp. 9–32). New York, NY: Routledge.

Zirkel, P. A., & Thomas, L. B. (2010). State laws for RTI: An updated snapshot. *Teaching Exceptional Children, 42(*3), 56–63. https://doi. org/10.1177/004005991004200306

Table 1. *Literature Review: Search Terms Used for Literature Search on Electronic Databases*

Category	Key words
Curriculum-based measurement	"curriculum based measure*" OR "curriculum-based measure*" OR "curriculum based measurement*" OR "curriculum-based measurement*" OR CBM OR "general outcome measure*" OR "progress monitoring" OR "mastery measure*" OR "curriculum based assessment*" OR "curriculum-based assessment*"
Writing	"correct letters sequences" OR CLS OR "correct minus incorrect letter sequences" OR CILS OR "correct word sequences" OR CWS OR "correct minus incorrect word sequences" OR CIWS OR handwriting OR keyboard* OR "picture word" OR "sentence construction" OR spelling OR "story prompt" OR transcription OR typing OR "word dictation" OR "words written" OR WW OR "words spelled correctly" OR WSC OR writing OR "writing fluency" OR "written expression" OR "written language*"
Multilingual students	bilingual* OR multilingual* OR "heritage language*" OR "home language*" OR "dual language*" OR "English learner*" OR "English language learner*" OR ELL OR "English as an additional language" OR "English as a second language" OR "immersion classroom*" OR "non-native*" OR "second language learner*" OR "multicultural class*" OR "L2-learner*" OR "dual language learner*" OR "multilingual learner*"

Table 2. *Literature Review: Coding Descriptors*

Descriptor	Definition
Multilingual Students	
Number of students	Sample size of multilingual students administered with English writing CBMs
Grade	Grade of multilingual students administered with English writing CBMs
Home languages (L1s)	Home languages of multilingual students as described by authors
Language instruction	Language of instruction at school or type of instruction (e.g., dual-language program, English as a Second Language classroom) the multilingual students were involved, as described by authors
Language proficiency	Language proficiency, either L1 or L2 (English), of the multilingual students, as described by authors
CBM	
Task/prompt type	Type of English writing CBM tasks (e.g., word dictation, passage copying) and/or prompts (e.g., narrative prompt, picture prompt) as described by authors
Writing level	Level of writing skills (e.g., word, sentence, discourse) targeted by English writing CBMs as described by authors
Time (min)	Response time (number of minutes) allowed for the students
Scoring	Scoring procedures (e.g., words written, correct word sequences) used for English writing CBMs
Validity evidence	
Criterion measures	Name of any criterion measures involved in calculating correlations with English writing CBMs

Note. CBM = curriculum-based measurement.

Table 3. *Literature Review: Rubric for Quality Indicators Rating*

Quality indicator	0 (Not met)	1 (Partially met)	2 (Met)
Description of the participant and instructional context			
Describe the participants' language background	Language background of participants not provided, or described only as "bi/multilingual" or "ELL" without home languages specified	Language background of participants partially described, without L1 and L2 language proficiency specified	Language background of participants provided, including level of proficiency in both L1 and L2
Describe the language instruction provided	Language instruction not specified, or simply described as students being eligible for ELL services	Language instruction partially described, but either time or content of instruction not specified	Language instruction described, with time and content of instruction provided
Description of Measures			
Provide evidence, from a prior study or test manual, that suggests validity evidence and other psychometric properties for English writing CBM of interest	Psychometric properties of English writing CBM not provided	Validity and/or other psychometric properties of English writing CBM partially described, without providing specific estimates	Validity evidence and/or other psychometric properties of English writing CBM explicitly described with exact estimates
Provide evidence, from a prior study or test manual, that suggests validity evidence and other psychometric properties	Psychometric properties of criterion measure not provided	Validity and/or other psychometric properties of criterion measure partially described, without providing specific estimates	Validity evidence and/or other psychometric properties of criterion measure explicitly described with exact estimates

for the criterion measure used (if applicable)			
Measurement and scoring			
Describe the administration and scoring process for English writing CBM of interest (e.g., training; if applicable)	Neither administration nor scoring process for English writing CBM described	Administration and scoring process for English writing CBM partially described	Administration and scoring process for English writing CBM described explicitly and sufficiently, including information about characteristics of administrators and training process for administration or scoring
Describe the administration and scoring process for criterion measure (if applicable)	Neither administration nor scoring process for criterion measure described	Administration and scoring process for criterion measure partially described	Administration and scoring process for criterion measure described explicitly and sufficiently, including information about characteristics of administrators and training process for administration or scoring
Attain and report sufficient interscorer reliability and/or fidelity of administration for English writing CBM	Neither interscorer reliability nor fidelity of administration measured for English writing CBM	Interscorer reliability and/or fidelity of administration measured for English writing CBM and partially described	Interscorer reliability and/or fidelity of administration measured and reported for English writing CBM, achieving 80% or above
Set and report a reasonable time period between English writing CBM of interest and criterion measure to ensure that the participant condition	Timing of administering English writing CBM and criterion measure not reported	Timing of administering English writing CBM and criterion measure partially described, without providing a rationale for the time period	Timing of administering English writing CBM and criterion measure described, with sufficient rationale supporting the appropriateness of such time interval

did not change between the two (if applicable)		between the two administrations	
Data analysis and reporting of findings			
Discuss and address assumptions of the analysis	Assumptions of the analysis not examined and reported	At least one assumption of the analysis, such as linearity, independence, normality, or homoscedasticity, examined and partially described, such as the assumption being met	At least one assumption of the analysis, such as linearity, independence, normality, or homoscedasticity, examined and explicitly reported along with the procedure and results
Report reliability and/or validity estimates for all measured variables	Reliability and/or validity estimates not provided	Reliability and/or validity estimates provided for some of the variables examined	Reliability and/or validity estimates provided for all variables examined
Significance of findings			
Discuss the practical relevance of reliability and validity estimates	Practical relevance of the findings on reliability and validity estimates of English writing CBM not discussed	Practical relevance of the findings on reliability and validity estimates of English writing CBM partially discussed	Practical relevance of the findings on reliability and validity estimates of English writing CBM sufficiently discussed

Note. CBM = curriculum-based measurement. ELL = English language learner.

Table 4. *Literature Review: Quality of Studies Included in Review*

	0 (not met)	1 (partially)	2 (met)
Description of the participant and instructional context			
Describe the participants' language background	0	7	0
Describe the language instruction provided	0	7	0
Description of Measures			
Provide evidence, from a prior study or test manual, that suggests validity evidence and other psychometric properties for English writing CBM of interest	2	0	5
Provide evidence, from a prior study or test manual, that suggests validity evidence and other psychometric properties for the criterion measure used (if applicable)	1	2	4
Measurement and scoring			
Describe the administration and scoring process for English writing CBM of interest (e.g., training; if applicable)	0	6	1
Describe the administration and scoring process for criterion measure (if applicable)	0	6	1
Attain and report sufficient interscorer reliability and/or fidelity of administration for English writing CBM	0	0	7
Set and report a reasonable time period between English writing CBM of interest and criterion measure to ensure that the participant condition did not change between the two (if applicable)	1	5	1
Data analysis and reporting of findings			
Discuss and address assumptions of the analysis	4	0	3
Report reliability and/or validity estimates for all measured variables	0	0	7
Significance of findings			
Discuss the practical relevance of reliability and validity estimates	0	0	7

Note. CBM = curriculum-based measurement.

Table 5. *Literature Review: Characteristics of the Included Studies*

	Multilingual students					English writing CBM			Criterion validity		Diagnostic validity	Reliability	
Study	*n*	Grade (s)	L1	Instruction	Proficiency	Prompt	Time (min)	Scoring	Criterion measure	Correlation		Alternate-form	Inte-scorer
Spanish as the L1 (with the highest percentage); Elementary grade level studies													
Smith & Lembke (2021)	73	1, 2, 3	Spanish (38.4%), Arabic (9.6%)	English-only	L1 NR; English Beginning to Advance	Picture word	3	WW WSC CWS IWS %WSC %CWS CIWS CWSR	ACCESS writing[a]	.43–.61 .42–.62 .38–.65 – -.01–.32 .06–.34 .09–.49 -.10–.56		.77–.97 .71–.96 .60–.94 .48–.91 .31–.73 .42–.86 .30–.91 .30–.94	88–98%
Smith & Lembke (2022)	73	1, 2, 3	Spanish (38.4%), Arabic (9.6%)	English-only	L1 NR; English Beginning to Advance	Word dictation	3	WW WSC CLS ILS %WSC %CLS CILS	ACCESS writing[b]	.26–.63 .57–.81 .56–.78 – .52–.70 .47–.60 .51–.75		.81–.97 .88–.97 .91–.97 .61–.98 .79–.95 .78–.93 .86–.97	94–99%
Keller-Margulis et al. (2016)	19	4	Languages NR; Asian (5.3%), Black (5.3%), Hispanic (89.5%)	Transitional bilingual; 75% instruction in English	NR	Story starter (AIMSweb)	3	TWW CWS WSC CP %CWS %WSC CIWS	STAAR	.02–.26 .08–.43 .05–.27 -.04–.31 -.02–.67 -.02–.33 .00–.68	SE=1.0(%CWS), SP=.73(%CWS), AUC=.76(CIWS)–.84(%CWS)		

	31	4	Languages NR; Asian (6.5%), Hispanic (93.6%)		NR					-.05–.02 .17–.20 -.00–.01 .24–.50 .08–.46 -.06–.36 .06–.38	SE=1.0 (%CWS,%WSC,CIWS), SP=.86(%WSC)–.93(CIWS), AUC=.47(CP)–.93(CIWS)		
Landis (2019)	175	2	Spanish	Transitional bilingual or two-way immersion	NR	Descriptive prompt	5	NDW CTTR	ELPA21 writing[c]	.24–.48 .20–.49	AUC .60–.81 AUC .55–.80	.62–.72 .51–.65	ICC = .96
Somali as the L1 (with the highest percentage); Secondary grade level studies													
Campbell (2010)	57	9-12	Somali (82.5%), Amharic (1.8%), Eritrean (1.8%), Oromo (5.3%), Spanish (3.5%), Tibetan (1.8%)	ESL	NR	Passage copying	1.5	WW CW CIW %CW CWS CIWS %CWS	Teacher ratings, MBST TEAE, TOWL-3	.35–.49 .38–.53 .33–.64 – .43–.67 .40–.62 –		.59–.67 .63–.73 .65–.69 .38–.48 .69–.72 .70 .51–.60	
Espin et al. (2008)	38	10	Somali (50%), Spanish (23%), Hmong (13%), Laotian (7%),	ESL	L1 NR; English Levels 4 and 5 ELL (fluent)	Narrative prompt	3, 5, 7, 10	WW WWC CWS CIWS	MBST/MCA	.33–.39 .37–.43 .56–.64 .70–.75		.55–.82 .59–.83 .61–.87 .64–.88	

			Yoruba Nigerian (2-3%), Chinese (2-3%), Oromo Ethiopian (2-3%)								
African Language (not specified) as the L1 (with the highest percentage); Secondary grade level studies											
Campbell et al. (2013)	36	10-12	African language 92% (Amharic, Eritrean, Oromo, Somali)	ESL	L1 NR; English Moderate to high	Narrative, expository, picture prompts	3, 5, 7	WW CW %CW CWS CIWS %CWS WW+CIWS #T-units Words/#T-unit	Teacher ratings, MBST TEAE, TOWL-3	.19–.66 .25–.72 .45–.82 .38–.82 .51–.84 .50–.80 .41–.83 .05–.58 -.13–.52	.56–.89 .58–.90 .02–.77 .60–.91 .49–.89 .45–.88 .59–.90 .53–.82 .13–.62

Note. ELL = English Language Learner. L1 = first language or home language. ESL = English as a Second Language. NR = not reported; #T-units = Number of T-units. CILS = Correct minus Incorrect Letter Sequences. CIW = Correct minus Incorrect Words. CIWS= Correct minus Incorrect Word Sequences. CIW = Correct minus Incorrect Words. CLS = Correct Letter Sequences. CP = Correct Punctuation. CTTR = Corrected Type Token Ratio. CW = Correct Words. CWS = Correct Word Sequences. CWSR = Correct Word Sequences per Response. ILS = Incorrect Letter Sequence. IWS = Incorrect Word Sequences. NDW = Number of Different Words. TWW = Total Words Written. WSC = Words Spelled Correctly. WW = Words Written. WWC = Word Written Correctly; ACCESS = Assessing Comprehension and Communication in English State-to-State English Language Proficiency test. ELPA21 = English Language Proficiency Assessment for the 21st Century. MBST = Minnesota Basic Skills Test (Minnesota Department of Education, 2005). MBST/MCA = Minnesota Basic Standards Test/Minnesota Comprehensive Assessments (Minnesota Department of Education & Beck Evaluation and Testing Associates, 1997). STAAR = State of Texas Assessments of Academic Readiness. TEAE =

Test of Emerging Academic English (Minnesota Department of Education, 2002). TOWL-3 = Test of Written Language-Third Edition (Hammill & Larsen, 1996); AUC = Area Under Curve. ICC = Intraclass Correlation Coefficient. SE = Sensitivity. SP = Specificity. ICC = Intraclass Correlation Coefficient.

[a]Smith & Lembke (2021) examined predictive validity coefficients in relation to writing, reading, and oral language subtests of ACCESS. Correlations with the writing subtest were included in this table.

[b]Smith & Lembke (2022) examined predictive and concurrent validity evidence. Concurrent validity coefficients were reported in this table; Smith & Lembke (2022) examined the concurrent validity coefficients in relation to writing, reading, speaking, and listening subtests of ACCESS. Correlations with the writing subtest were included in this table.

[c]Landis (2019) examined correlational validity coefficients in relation to writing, reading, speaking, and listening subtests of ELPA21. Correlations with the writing subtest were included in this table. Similarly, ROC analysis results using the 20th percentile cut score in the writing subtest were reported in this table.

Table 6. *Literature Review: Findings of Reliability and Validity Evidence Based on Complexity, Accuracy, and Fluency Framework Classification*

	Fluency							Fluency + Accuracy				Accuracy			Complexity				
	WW	WSC	CWS	IWS	CP	CLS	ILS	CIW	CIWS	CILS	WW + CIWS	%CWS	%WSC	%CLS	NDW	CTTR	CWSR	T-units	W/T-units
Alternate-form reliability																			
Smtih & Lembke (2021)	.77–.97	.71–.96	.60–.94	.48–.91					.30–.91			.42–.86	.31–.73				.30–.94		
Smith & Lembke (2022)	.81–.97	.88–.97				.91–.97	.61–.98			.86–.97			.79–.95	.78–.93					
Landis (2019)															.62–.72	.51–.65			
Campbell (2010)	.59–.67	.63–.73	.69–.72					.65–.69	.70			.51–.60	.38–.48						
Espin et al. (2008)	.55–.82	.59–.83	.61–.87						.64–.88										
Campbell et al. (2013)	.56–.89	.58–.90	.60–.91						.49–.89		.59–.90	.45–.88	.02–.77					.53–.82	.13–.62
Criterion validity																			
Smith & Lembk	.43–.61	.42–.62	.38–.65						.09–.49			.06–.34	-.01–.32				-.10–.56		

e (2021)																			
Smith & Lembke (2022)	.26–.63	.57–.81				.56–.78				.51–.75			.52–.70	.47–.60					
Keller-Margulis et al. (2016), Sample 1	.02–.26	.05–.27	.08–.43		-.04–.31				.00–.68			-.02–.67	-.02–.33						
Keller-Margulis et al. (2016), Sample 2	-.05–.02	-.00–.01	.17–.20		.24–.50				.06–.38			.08–.46	-.06–.36						
Landis (2019)															.24–.48	.20–.49			
Campbell (2010)	.35–.49	.38–.53	.43–.67					.33–.64	.40–.62										
Espin et al. (2008)	.33–.39	.37–.43	.56–.64						.70–.75										
Campbell et al. (2013)	.19–.66	.25–.72	.38–.82						.51–.84		.41–.83	.50–.80	.45–.82					.05–.58	-.13–.52
Diagnostic validity																			

Keller-Margulis et al. (2016), Sample 1		SE 1.0, SP .73, AUC .84			
Keller-Margulis et al. (2016), Sample 2	SE 1.0, SP .93, AUC . 93	SE 1.0, SP .90, AUC .90	SE 1.0, SP .86, AUC . 86		
Landis (2019)				AUC .60–.81	AUC . 55 –.80

Note. The reported reliability and criterion validity values represent Pearson's *r* correlations; WW = Words Written, WSC = Words Spelled Correctly, CWS = Correct Word Sequences, IWS = Incorrect Word Sequences, CP = Correct Punctuation, CLS = Correct Letter Sequences, ILS = Incorrect Letter Sequences, CIW = Correct minus Incorrect Words, CIWS = Correct minus Incorrect Word Sequence, CILS = Correct minus Incorrect Letter Sequences, NDW= Number of Different Words, CTTR = Corrected Type Token Ratio, CWSR = Correct Word Sequences per Response, #T-units = Number of T-units, W/T-units = Words/T-units, SE = Sensitivity, SP = Specificity, AUC = Area Under Curve.

Table 7. *Students' Demographics by Language Condition*

	Multilingual, *n* (%)	English-monolingual, *n* (%)	$\chi^2(p)$
Grade			17.47 (.002)
K-1	11 (16.4%)	61 (21.6%)	
2	32 (47.8%)	74 (26.2%)	
3	11 (16.4%)	88 (31.2%)	
4	7 (10.4%)	48 (17.0%)	
5-6	6 (9.0%)	11 (3.9%)	
Sex			0.60 (.741)
Female	18 (26.9%)	86 (30.5%)	
Male	49 (73.1%)	195 (69.1%)	
No response	0 (0%)	1 (0.4%)	
Race/ethnicity			154.49 (<.001)
Native American/Alaskan	1 (1.5%)	3 (1.1%)	
Black/African American	16 (23.9%)	46 (16.3%)	
Asian American/Pacific Islander	10 (14.9%)	4 (1.4%)	
Hispanic/Latino(a) American	34 (50.7%)	13 (4.6%)	
White/European American	4 (6.0%)	198 (70.2%)	
Not listed	0 (0%)	2 (0.7%)	
Multiracial	1 (1.5%)	15 (5.3%)	
Other	1 (1.5%)	1 (0.4%)	
Special education			42.39 (<.001)
Eligible	43 (64.2%)	263 (93.3%)	
Not eligible	24 (35.8%)	19 (6.7%)	
Primary disability category			24.66 (.017)

Autism	14 (32.6%)	56 (21.3%)	
Emotional/behavior disorder	0 (0%)	25 (9.5%)	
Physical impairment	0 (0%)	1 (0.4%)	
Deaf/blind	0 (0%)	2 (0.8%)	
Severely multiply impaired	0 (0%)	1 (0.4%)	
Deaf/hard of hearing	3 (7.0%)	4 (1.5%)	
Intellectual disability	4 (9.3%)	15 (5.7%)	
Other health disability	6 (14.0%)	66 (25.1%)	
Specific learning disability	8 (18.6%)	64 (24.3%)	
Traumatic brain injury	0 (0%)	1 (0.4%)	
Need alternative programming	6 (14.0%)	9 (3.4%)	
Developmental delay	1 (2.3%)	5 (1.9%)	
Speech/language impairment	1 (2.3%)	13 (4.9%)	
No response	0	1 (0.4%)	
English Learner service			258.75 (<.001)
Eligible	55 (82.1%)	4 (1.4%)	
Not eligible	5 (7.5%)	255 (90.4%)	
No response	7 (10.4%)	23 (8.2%)	
Home languages			349 (<.001)
English	0 (0%)	282 (100.0%)	
Spanish	33 (49.3%)	0 (0%)	
Somali	14 (20.9%)	0 (0%)	
Hmong	1 (1.5%)	0 (0%)	
Other	19 (28.4%)	0 (0%)	

Note. Under Grade, K-1 sample includes $n = 1$ kindergartener, and the 5-6 sample includes $n = 2$ 6th-graders.

Table 8. *Comparing Multilingual and English-Monolingual Students' Writing Skills as Measured by Different Writing Assessments*

	Multilingual (n = 67), M (SD)	English-monolingual (n = 282), M (SD)	t(p)	Hedges' g
Picture Word (correct word sequences)				
Form A	8.05 (7.30)	7.88 (8.44)	-.158 (.875)	.02
Form B	7.58 (7.55)	8.09 (9.19)	.463 (.644)	-.06
Story Prompt (correct word sequences)				
Form A	4.63 (4.92)	4.55 (5.59)	.289 (.774)	-.04
Form B	5.53 (6.45)	3.66 (5.02)	.573 (.568)	-.08
KTEA-3	71.85 (16.42)	68.24 (13.47)	-.459 (.648)	.07

Note. KTEA = Kaufman Test of Educational Achievement-Third Edition (Kaufman & Kaufman, 2014). For KTEA-3, Written Language composite scores (based on Written Expression and Spelling subtests) were used.

Table 9. *Word Lists for Word Dictation Forms A and B*

Item	Form A	Form B	Item	Form A	Form B
1	**CUT**	**LIP**	16	HOPE	ROBE
2	**SWAP**	**TRIM**	17	STOVE	GRAPE
3	**SACK**	**LEND**	18	LOAF	POOL
4	**PAT**	**ROB**	19	PLOT	PLUS
5	**SKIT**	**PROP**	20	JUMP	LAND
6	**MELT**	**LIFT**	21	LAKE	KITE
7	**PAGE**	**FINE**	22	CHEAP	SNEAK
8	**PET**	**RUG**	23	FROST	CRUSH
9	**FLAP**	**CHIN**	24	PRIDE	SNORE
10	**SILK**	**KEPT**	25	COOL	GAIN
11	**MILE**	**TIRE**	26	RAKE	POLE
12	**STORE**	**TRADE**	27	TRAIL	CLEAR
13	**SEED/CEDE**	**MOAN**	28	SHOCK	STAND
14	**TWIN**	**CROP**	29	STONE	PLANE
15	**KIND**	BULB	30	TOAD	MOOD

Note. Bolded words were included for analysis.

Table 10. *Descriptive Statistics for Assessment-Level Scores by Language Condition*

	Entire (N = 349)					Multilingual (n = 67)					English-monolingual (n = 282)					Comparison	
	n	*M* (*SD*)	Range	Skw.	Kurt.	*n*	*M* (*SD*)	Range	Skw.	Kurt.	*n*	*M* (*SD*)	Range	Skw.	Kurt.	*t*(*p*)	Hedges' *g*
Form A																	
WW	343	13.88 (7.92)	[0, 43]	0.22	-0.11	66	13.03 (7.05)	[0, 32]	0.42	-0.22	277	14.09 (8.11)	[0, 43]	0.17	-0.13	1.061 (.291)	.00
WSC	343	4.20 (5.70)	[0, 39]	2.54	8.60	66	4.48 (6.12)	[0, 30]	2.01	4.20	277	4.13 (5.61)	[0, 39]	2.68	9.91	-.430 (0.668)	.00
CLS	343	39.83 (33.10)	[0, 213]	1.20	2.31	66	38.92 (33.88)	[1, 164]	1.33	1.70	277	39.97 (32.97)	[0, 213]	1.16	2.44	.242 (.809)	.00
Form B																	
WW	335	14.67 (8.49)	[0, 41]	0.39	0.29	66	13.76 (7.76)	[0, 40]	0.59	0.71	269	14.90 (8.69)	[0, 41]	0.33	0.19	1.444 (.152)	.00
WSC	335	4.05 (5.43)	[0, 38]	2.58	10.06	66	4.27 (6.38)	[0, 37]	2.73	9.69	269	3.99 (5.19)	[0, 38]	2.43	9.34	-.265 (.791)	.00
CLS	335	39.18 (34.43)	[0, 203]	1.39	2.88	66	37.79 (36.62)	[0, 200]	1.77	4.52	269	39.52 (33.94)	[0, 203]	1.26	2.29	.490 (.625)	.00

Note. WW = words written; WSC = words spelled correctly; CLS = correct letter sequences; Skw. = skewness; Kurt. = kurtosis. Students who did not provide any response to any of the 30 items in each form were excluded.
If a student finished all 30 words before the 3-minute time limit, their scores were adjusted proportionately based on their completion time, leading to total scores exceeding 30 for WW and WSC.

Table 11. *Descriptive Statistics of Item-Level Scores by Language Condition*

	Entire (N = 349)					Multilingual (n = 67)					English-monolingual (n = 282)				
	Missing	WSC		CLS		Missing	WSC		CLS		Missing	WSC		CLS	
		M (*SD*)	ITC	*M* (*SD*)	ITC		*M* (*SD*)	ITC	*M* (*SD*)	ITC		*M* (*SD*)	ITC	*M* (*SD*)	ITC
Form A *(Items 16-30 excluded)*															
1 (CUT)	2.9%	.48 (.50)	.58	2.54 (1.55)	.69	3.0%	.42 (.50)	.66	2.45 (1.49)	.67	2.8%	.49 (.50)	.56	2.57 (1.56)	.69
2 (SWAP)	4.7%	.08 (.27)	.54	1.96 (1.35)	.74	6.1%	.15 (.36)	.67	2.24 (1.51)	.72	4.3%	.06 (.24)	.50	1.89 (1.30)	.75
3 (SACK)	5.5%	.12 (.33)	.58	2.33 (1.47)	.77	3.0%	.08 (.27)	.53	2.22 (1.37)	.69	6.1%	.13 (.34)	.60	2.35 (1.50)	.79
4 (PAT)	4.1%	.60 (.49)	.56	2.88 (1.50)	.72	1.5%	.49 (.50)	.52	2.62 (1.52)	.58	4.7%	.62 (.49)	.58	2.94 (1.50)	.75
5 (SKIT)	7.9%	.18 (.39)	.65	2.46 (1.55)	.80	6.1%	.24 (.43)	.76	2.48 (1.64)	.81	8.3%	.17 (.37)	.62	2.45 (1.53)	.80
6 (MELT)	11.1%	.24 (.43)	.73	2.72 (1.58)	.79	9.1%	.25 (.44)	.70	2.85 (1.55)	.70	11.6%	.24 (.43)	.74	2.69 (1.58)	.81
7 (PAGE)	12.2%	.18 (.38)	.66	2.24 (1.58)	.72	10.6%	.25 (.44)	.76	2.42 (1.71)	.79	12.6%	.16 (.36)	.63	2.20 (1.55)	.71
8 (PET)	13.4%	.52 (.50)	.61	2.76 (1.39)	.71	15.2%	.54 (.50)	.64	2.86 (1.30)	.71	13.0%	.51 (.50)	.61	2.74 (1.42)	.71

9 (FLAP)	16.6%	.36 (.48)	.65	2.87 (1.86)	.80	18.2%	.43 (.50)	.54	3.19 (1.79)	.66	16.2%	.35 (.48)	.68	2.80 (1.87)	.83
10 (SILK)	21.0%	.13 (.34)	.53	2.21 (1.54)	.78	25.8%	.27 (.45)	.62	2.64 (1.69)	.80	19.9%	.10 (.31)	.50	2.11 (1.49)	.78
11 (MILE)	25.1%	.17 (.37)	.71	2.48 (1.51)	.78	31.8%	.22 (.42)	.82	2.62 (1.63)	.85	23.5%	.16 (.36)	.67	2.45 (1.49)	.76
12 (STORE)	31.2%	.13 (.33)	.66	2.71 (1.80)	.77	40.9%	.21 (.41)	.78	3.18 (1.90)	.79	28.9%	.11 (.32)	.62	2.61 (1.77)	.77
13 (SEED/CEDE)	35.3%	.24 (.43)	.62	2.70 (1.60)	.75	45.5%	.25 (.44)	.71	2.78 (1.51)	.77	32.9%	.24 (.43)	.61	2.68 (1.62)	.74
14 (TWIN)	42.9%	.30 (.46)	.66	2.83 (1.65)	.77	54.5%	.30 (.47)	.72	3.00 (1.58)	.76	40.1%	.30 (.46)	.65	2.80 (1.66)	.77
15 (KIND)	49.0%	.35 (.48)	.65	2.84 (1.86)	.77	62.1%	.56 (.51)	.72	3.52 (1.90)	.80	45.8%	.31 (.47)	.62	2.73 (1.84)	.76
Form B *(Items 15-30 excluded)*															
1 (LIP)	3.9%	.49 (.50)	.61	2.59 (1.51)	.75	3.0%	.48 (.50)	.68	2.62 (1.49)	.78	4.1%	.49 (.50)	.59	2.58 (1.52)	.75
2 (TRIM)	7.2%	.23 (.42)	.68	2.32 (1.81)	.83	7.6%	.26 (.44)	.63	2.33 (1.93)	.79	7.1%	.22 (.42)	.69	2.31 (1.79)	.83
3 (LEND)	4.5%	.14 (.34)	.60	2.06 (1.57)	.79	4.5%	.11 (.32)	.59	1.94 (1.54)	.78	4.5%	.14 (.35)	.60	2.09 (1.58)	.80
4 (ROB)	3.9%	.39 (.49)	.54	2.32 (1.54)	.70	3.0%	.33 (.47)	.56	2.20 (1.48)	.74	4.1%	.40 (.49)	.54	2.35 (1.56)	.69

5 (PROP)	7.2%	.36 (.48)	.76	2.77 (1.93)	.83	4.5%	.30 (.46)	.63	2.70 (1.86)	.76	7.8%	.38 (.49)	.80	2.78 (1.95)	.84
6 (LIFT)	6.9%	.31 (.46)	.72	2.77 (1.78)	.84	7.6%	.33 (.47)	.80	2.79 (1.83)	.88	6.7%	.30 (.46)	.69	2.76 (1.78)	.83
7 (FINE)	9.6%	.18 (.39)	.62	2.43 (1.61)	.80	9.1%	.22 (.42)	.57	2.35 (1.71)	.72	9.7%	.18 (.38)	.64	2.45 (1.59)	.82
8 (RUG)	11.3%	.40 (.49)	.68	2.28 (1.54)	.76	16.7%	.35 (.48)	.74	2.10 (1.49)	.77	10.0%	.42 (.49)	.67	2.32 (1.54)	.75
9 (CHIN)	17.0%	.32 (.47)	.68	2.57 (2.01)	.80	24.2%	.38 (.49)	.76	2.90 (1.98)	.85	15.2%	.31 (.46)	.67	2.50 (2.01)	.79
10 (KEPT)	19.1%	.09 (.28)	.42	1.63 (1.51)	.62	28.8%	.19 (.40)	.60	2.23 (1.77)	.65	16.7%	.07 (.25)	.36	1.51 (1.42)	.63
11 (TIRE)	21.8%	.17 (.37)	.57	2.34 (1.56)	.73	33.3%	.20 (.41)	.49	2.39 (1.60)	.58	19.0%	.16 (.37)	.59	2.33 (1.56)	.76
12 (TRADE)	26.9%	.09 (.28)	.54	2.17 (1.85)	.79	36.4%	.14 (.35)	.65	2.57 (2.09)	.84	24.5%	.07 (.26)	.51	2.08 (1.80)	.78
13 (MOAN)	31.9%	.04 (.18)	.40	2.07 (1.15)	.67	40.9%	.05 (.22)	.54	2.22 (1.17)	.58	29.7%	.03 (.18)	.36	2.03 (1.15)	.69
14 (CROP)	35.5%	.31 (.47)	.64	2.77 (1.87)	.81	47.0%	.37 (.49)	.62	3.15 (1.76)	.80	32.7%	.30 (.46)	.64	2.70 (1.89)	.80

Note. WSC = words spelled correctly; CLS = correct letter sequences; ITC = item-total correlation (Pearson's *r*; pairwise complete correlation was used).
Students who did not provide any response to any of the 30 items in each form were excluded from the analysis.

Table 12. *Evaluation of Missingness in Item-Level Data*

		Item 15 in Form A			Item 14 in Form A		
Variable	Subcategories	Not missing %	Missing %	*p*	Not missing %	Missing %	*p*
Grade	K-1	21.4	78.6	< .001	39.4	60.6	< .001
	2	43.7	56.3		58.8	41.2	
	3	65.3	34.7		75.0	25.0	
	4	72.7	27.3		83.3	16.7	
	5	64.7	35.3		76.5	23.5	
Sex	Male	49.8	50.2	.510	63.4	36.6	.648
	Female	54.4	45.6		66.7	33.3	
Special education eligibility	Not eligible	53.5	46.5	.855	62.8	37.2	.939
	Eligible	50.7	49.3		64.7	35.3	
Language	English-monolingual	54.2	45.8	.025	67.3	32.7	.043
	Multilingual	37.9	62.1		53.0	47.0	

Note. Variables (all categorical) were compared using a chi-squared test.

Table 13. *Q3 Values for Item Pairs, Scored Based on Words Spelled Correctly*

Form A (Full Sample, *N* = 343)

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	-.186													
3	-.021	-.141												
4	.104	-.009	.004											
5	-.104	.007	-.092	-.042										
6	-.094	.017	-.198	.006	-.094									
7	.016	-.028	.055	-.078	.034	**-.244**								
8	.049	-.098	.005	.047	.022	-.041	-.099							
9	.020	.137	.054	-.083	-.086	-.074	.010	**-.201**						
10	-.132	**-.254**	.093	-.066	**.284**	-.126	-.028	-.024	-.217					
11	.026	-.101	-.190	-.027	**-.239**	-.210	-.100	-.170	-.106	.066				
12	-.050	-.102	.048	-.081	-.210	-.103	-.062	-.075	-.098	-.136	-.186			
13	.066	-.189	-.002	.001	-.129	.003	-.137	.067	-.055	-.043	-.127	-.064		
14	.019	-.108	-.084	.001	-.022	-.137	-.233	-.113	-.008	**-.213**	.178	-.036	-.093	
15	-.120	.129	-.170	-.196	-.079	-.024	-.023	-.073	-.116	.099	-.091	-.080	-.061	**-.223**

Average = -.062
SD = .098
Pairs with local dependence = 7 (6.7%)

Form B (Full Sample, *N* = 335)

Item	1	2	3	4	5	6	7	8	9	10	11	12	13
2	.022												
3	-.064	-.104											
4	.107	-.030	-.001										
5	-.014	-.119	-.039	-.089									
6	-.086	-.029	-.010	-.118	-.157								
7	**-.254**	**-.264**	.035	-.131	-.107	-.167							
8	-.005	-.111	-.154	.010	-.146	-.172	-.093						
9	-.072	-.025	-.055	-.098	**-.285**	.066	-.146	-.031					
10	-.018	-.055	-.088	-.073	-.043	**-.236**	.052	.158	-.029				
11	-.019	-.180	-.025	-.154	-.025	.005	**.205**	**-.200**	-.134	-.063			
12	-.076	-.149	-.114	-.096	-.088	-.128	-.058	.093	-.180	.026	-.035		
13	.039	-.089	-.173	.102	-.083	-.148	.067	.033	-.139	-.034	-.031	.033	
14	-.043	.025	-.108	.063	-.079	-.086	-.021	**-.263**	.004	**-.201**	-.140	-.099	-.070

Average = -.067
SD = .092
Pairs with local dependence = 8 (8.8%)

Note. WSC = words spelled correctly.
The $|Q_3|$ values above the cut-off (.20), indicative of local dependency, appear in bold face.
Students who did not provide any response to any of the 30 items in each form were excluded from the analysis.

Table 14. *Jackknife Slope Index (JSI) Values for Item Pairs, Scored Based on Words Spelled Correctly*

Form A (Full Sample, *N* = 343)

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	-.189													
3	.121	-.226												
4	.433	.005	-.119											
5	-.125	.147	-.081	-.129										
6	-.152	.345	-.359	.236	.100									
7	.195	.144	.230	-.120	.323	-.528								
8	.200	-.216	-.020	.456	.068	.317	-.156							
9	.321	.556	.137	-.219	.230	.215	.405	-.605						
10	-.113	-.262	.357	-.136	**.732**	-.087	.169	-.076	-.184					
11	.142	-.030	-.291	-.091	-.389	-.204	.098	-.580	.255	**.616**				
12	-.051	.206	.354	-.180	-.344	.046	.223	-.235	.028	-.008	.083			
13	.187	-.275	-.018	.128	-.031	.238	-.177	.468	.126	.012	-.198	-.001		
14	.114	-.106	-.144	-.006	.016	-.018	-.244	-.466	.581	-.242	**1.085**	.202	-.146	
15	.137	.574	-.095	-.393	-.028	.253	.221	-.079	.114	.323	.057	-.058	-.080	-.508

Average = .033

SD = .284
Pairs with local dependence = 3 (2.9%)

Form B (Full Sample, *N* = 335)

Item	1	2	3	4	5	6	7	8	9	10	11	12	13
2	.257												
3	-.131	-.101											
4	.265	.105	-.020										
5	.266	.031	.272	.031									
6	-.041	.393	.390	-.075	.059								
7	-.610	-.603	.335	-.245	.151	-.239							
8	.234	.212	-.283	.351	-.030	.078	-.039						
9	.038	.463	.059	-.067	-.646	**.927**	-.231	.421					
10	-.007	.020	-.093	-.047	.068	-.437	.491	.638	.019				
11	-.274	-.269	.177	-.291	.450	.390	**.998**	-.311	-.119	.048			
12	-.131	-.127	-.028	-.115	.002	-.165	.364	.658	-.370	.410	.250		
13	.056	.067	-.285	.259	-.053	-.357	.495	.265	-.354	.257	.024	**.709**	
14	-.023	.496	-.073	.364	.383	.148	.197	-.424	.334	-.285	-.081	.029	.086

Average = .070

SD = .316
Pairs with local dependence = 3 (3.3%)

Note. WSC = words spelled correctly.
The mean of the JSI values (folded/summed) plus twice the standard deviation was used as the cutoff value. The JSI values above the cut-off, indicative of local dependency, appear in bold face.
Students who did not provide any response to any of the 30 items in each form were excluded from the analysis.

Table 15. *Q3 Values for Item Pairs, Scored Based on Correct Letter Sequences*

Form A (Full Sample, $N = 343$)

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	.035													
3	.014	-.070												
4	.017	.079	.000											
5	-.152	.019	-.052	-.056										
6	-.095	-.031	-.142	.013	-.145									
7	-.106	-.025	.076	-.025	-.073	-.177								
8	.131	-.117	**-.224**	.034	-.134	.052	-.088							
9	-.082	.083	.025	-.154	-.153	-.160	.019	-.105						
10	**-.204**	**-.256**	.053	-.062	.134	-.121	-.103	-.120	-.135					
11	.032	-.199	**-.246**	-.098	-.068	-.015	-.055	-.024	-.172	-.032				
12	-.087	-.097	-.063	-.120	-.083	-.148	-.021	-.120	-.013	-.042	-.021			
13	.122	-.140	.036	-.038	-.148	-.004	-.035	-.044	-.060	.007	-.183	-.181		
14	.059	-.090	-.153	-.030	.032	.004	**-.230**	-.045	-.183	-.151	.048	.059	-.162	
15	**-.210**	-.041	-.144	-.152	-.098	-.003	-.045	-.113	-.039	.046	.001	-.009	.022	-.184

Average = -.064
SD = .088
Pairs with local dependence = 6 (5.7%)

Form B (Full Sample, *N* = 335)

Item	1	2	3	4	5	6	7	8	9	10	11	12	13
2	-.040												
3	-.166	-.178											
4	.018	-.073	-.049										
5	-.052	-.117	-.065	-.151									
6	-.083	-.093	-.042	-.032	-.160								
7	-.150	-.113	-.003	-.072	-.131	-.120							
8	.029	-.030	-.129	.095	-.039	-.092	-.030						
9	-.121	-.129	-.034	-.065	-.147	-.025	-.065	-.118					
10	-.076	-.124	-.096	-.070	-.013	-.130	.045	-.005	.031				
11	-.174	-.085	.025	-.164	-.105	.003	.122	-.163	-.166	-.090			
12	.013	.144	-.132	-.194	-.090	-.077	-.082	-.068	**-.213**	.027	.072		
13	.034	-.048	-.145	-.023	.021	-.090	-.117	-.011	.059	-.048	-.164	**-.262**	
14	.033	-.099	-.096	.033	.115	-.148	-.190	-.149	-.062	-.063	-.172	-.089	-.032

Average = -.070
SD = .080
Pairs with local dependence = 2 (2.2%)

Note. CLS = correct letter sequences.
The $|Q_3|$ values above the cut-off (.20), indicative of local dependency, appear in bold face.
Students who did not provide any response to any of the 30 items in each form were excluded from the analysis.

Table 16. *Jackknife Slope Index (JSI) Values for Item Pairs, Scored Based on Correct Letter Sequences*

Form A (Full Sample, *N* = 343)

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	.108													
3	.031	.022												
4	.141	.201	.146											
5	-.157	.217	.092	.077										
6	-.037	.114	-.317	.147	-.146									
7	.033	.103	.284	.010	-.090	-.126								
8	.324	.026	-.449	.228	-.175	**.587**	.093							
9	-.036	**.650**	.327	-.091	-.279	-.200	.627	.077						
10	-.324	-.360	.477	-.193	.780	-.505	-.055	-.404	-.388					
11	.270	-.267	-.423	-.066	.047	.406	.030	.245	-.237	-.053				
12	.032	.127	.306	-.124	.074	-.227	.176	-.223	.153	.054	.114			
13	.286	-.153	.360	-.005	-.193	.258	.134	-.079	.082	.274	-.331	-.067		
14	.245	-.009	-.123	.120	.176	.220	-.494	.198	-.497	-.214	.448	.250	-.172	
15	-.315	.055	-.395	-.164	-.043	.485	.072	-.013	.194	.358	.277	.050	.279	-.485

Average = .029

SD = .266
Pairs with local dependence = 2 (1.9%)

Form B (Full Sample, *N* = 335)

Item	1	2	3	4	5	6	7	8	9	10	11	12	13
2	.125												
3	-.099	**-.447**											
4	.268	-.029	.115										
5	.020	-.114	.098	-.156									
6	.032	.147	.123	.070	-.346								
7	-.341	-.139	.340	-.104	-.037	-.130							
8	.295	-.061	-.182	.408	.130	-.026	.097						
9	-.058	-.082	.185	.035	-.092	.286	-.068	-.065					
10	-.035	-.243	-.084	-.025	.208	-.071	.295	.037	.222				
11	-.054	.126	.232	-.150	-.079	.422	**.562**	-.152	-.032	.026			
12	.258	.709	-.175	-.293	.031	.050	.073	.058	-.440	.256	**.536**		
13	.300	.233	-.326	.242	.111	.012	-.253	.277	.433	.137	-.103	-.271	
14	.290	.102	-.157	.220	.443	-.295	-.266	-.283	.192	.037	-.304	.035	.240

Average = 0.039

SD = 0.229
Pairs with local dependence = 3 (3.3%)

Note. CLS = correct letter sequences.
The mean of the JSI values (folded/summed) plus twice the standard deviation was used as the cutoff value. The JSI values above the cut-off, indicative of local dependency, appear in bold face.
Students who did not provide any response to any of the 30 items in each form were excluded from the analysis.

Table 17. *Eigenvalues from Exploratory Factor Analysis Using Principal Components Estimation for Polytomous Items (Correct Letter Sequences)*

	Component 1		Component 2		Ratio Component 1/Component 2	
Statistic	Form A	Form B	Form A	Form B	Form A	Form B
Eigenvalues	19.528	23.057	1.952	2.020	10.0	11.4
Explained variance	55.8%	60.7%	5.8%	5.3%		

Table 18. *Global Goodness-of-Fit Statistics for Unidimensional Rasch Model and Partial Credit Model*

	M_2 or M_2*	*df*	*p*	CFI	TLI	RMSEA [95% CI]	SRMSR
Rasch model (Words Spelled Correctly)							
Form A	212.727	104	<.001	.945	.945	.079 [.064, .094]	.105
Form B	175.674	90	<.001	.953	.952	.070 [.055, .086]	.083
Partial credit model (Correct Letter Sequences)							
Form A	68.120	47	.024	.889	.887	.052 [.020, .138]	.138
Form B	45.202	39	.229	.923	.921	.029 [.000, .098]	.100

Note. CFI = comparative fit index; TLI = Tucker Lewis index; RMSEA = root mean square error of approximation; SRMSR = standardized root mean square residuals; CI = confidence interval. Students who did not provide any response to any of the 30 items in each form were excluded from the analysis.

Table 19. *Summary of Differential Item and Test Functioning Analyses*

	WSC											CLS			
	LRT		Raju's area		Raju's DFIT		Logistic regression		M-H		Effect Size (ESSD)	LRT		Raju's DFIT	Effect Size (ESSD)
	$\Delta G^2(df=1)$	Adj. p	Raju's Z	Adj. p	NCDIF (cutoff)	Power	ΔG^2	Adj. p	M-H χ^2	Adj. p		$\Delta G^2(df)$	Adj. p	NCDIF (cutoff)	
Form A															
1 (CUT)	2.384	.460	1.86	.232	.007 (.023)	.208	2.243	.303	1.287	.550	-	-	-	-	-
2 (SWAP)	1.443	.593	-1.482	.415	.003 (.005)	.277	3.341	.212	1.924	.492	-	-	-	-	-
3 (SACK)	4.829	.210	1.871	.232	**.003** (.002)	.586	4.784	.144	2.605	.492	-.417	7.923 (5)	.321	0 (.150)	0
4 (PAT)	6.019	.210	2.891	.058	**.031** (.021)	.646	5.564	.139	6.113	.201	-.498	12.684 (4)	.052	**.371** (.096)	-.399
5 (SKIT)	.309	.789	-.637	.714	.006 (.012)	.256	1.753	.348	.004	.951	-	-	-	-	-
6 (MELT)	.182	.837	.496	.715	.000 (.012)	.070	0.047	.887	.309	.723	-	-	-	-	-
7 (PAGE)	.993	.613	-1.082	.598	.011 (.012)	.427	3.272	.212	.016	.951	-	-	-	-	-
8 (PET)	.116	.847	.506	.715	.002 (.024)	.082	.020	.887	.095	.875	-	-	-	-	-
9 (FLAP)	.036	.882	-.191	.849	.008 (.021)	.270	.403	.657	.434	.695	-	-	-	-	-

10 (SILK)	3.136	.383	-2.058	.232	**.014** (.011)	.617	5.544	.139	1.665	.492	.325	6.524 (5)	.345	.023 (.150)	0
11 (MILE)	.022	.882	-.308	.813	.002 (.007)	.163	.246	.715	.616	.695	-	-	-	-	-
12 (STORE)	.465	.743	-.878	.648	.003 (.007)	.224	.959	.491	1.923	.492	-	-	-	-	-
13 (SEED/CEDE)	.564	.743	.668	.714	.000 (.009)	.046	.704	.548	.490	.695	-	-	-	-	-
14 (TWIN)	.962	.613	.862	.648	.001 (.009)	.102	1.547	.356	.445	.695	-	-	-	-	-
15 (KIND)	1.397	.593	-1.368	.428	**.037** (.019)	.787	2.162	.303	2.739	.492	.327	5.057 (5)	.409	**.188** (.150)	-.020
	ETSSD = -.017											ETSSD = -.027			
Form B															
1 (LIP)	.480	.997	.188	.916	.000 (.023)	.048	.100	.752	.002	.967	-	-	-	-	-
2 (TRIM)	.008	.997	-.637	.815	.003 (.013)	.168	1.942	.507	.507	.741	-	-	-	-	-
3 (LEND)	2.018	.435	.907	.815	.001 (.004)	.116	.830	.507	.172	.863	-	-	-	-	-
4 (ROB)	3.559	.364	1.347	.693	.010 (.016)	.336	1.216	.507	.965	.571	-	-	-	-	-
5 (PROP)	3.105	.364	1.287	.693	.008 (.017)	.241	1.120	.507	1.919	.465	-	-	-	-	-

6 (LIFT)	.081	.997	-.245	.916	.002 (.018)	.098	.933	.507	.345	.780	-	-	-	-	-
7 (FINE)	.044	.997	-.700	.815	.003 (.009)	.198	1.702	.507	.058	.945	-	-	-	-	-
8 (RUG)	3.522	.364	1.327	.693	.011 (.016)	.389	1.651	.507	.005	.967	-	-	-	-	-
9 (CHIN)	.002	.997	-.667	.815	.005 (.018)	.182	1.459	.507	2.306	.451	-	-	-	-	-
10 (KEPT)	2.543	.388	-2.233	.358	.005 (.006)	.435	5.250	.307	5.762	.229	-	-	-	-	-
11 (TIRE)	.282	.997	-.094	.925	.000 (.007)	.054	.037	.912	2.910	.451	-	-	-	-	-
12 (TRADE)	.202	.997	-.985	.815	.000 (.004)	.153	1.153	.507	2.685	.451	-	-	-	-	-
13 (MOAN)	.000	.997	-.340	.916	.000 (.001)	.068	.008	.928	1.239	.531	-	-	-	-	-
14 (CROP)	.053	.997	-.361	.916	.004 (.014)	.089	.227	.807	1.510	.511	-	-	-	-	-
	ETSSD = 0										ETSSD = 0				

Note. WSC = words spelled correctly; CLS = correct letter sequences; LRT = likelihood ratio test; NCDIF = non-compensatory differential item functioning; DFIT = differential functioning of items and tests (Raju et al., 1995); M-H = Mantel-Haenszel (Holland & Thayer, 1988); ESSD = expected score standardized difference; ETSSD = expected test score standardized difference; Adj. *p* = adjusted *p*-values (based on Benjamini-Hochberg's method).

DIF effect sizes were only calculated for items identified as having DIF by at least one method.

For polytomous item DIF analysis with CLS scores, items identified as not showing DIF by any of the methods used for dichotomous item DIF analyses were used as anchor items. Hence, DIF was not assessed for these anchor items.

Raju's *Z* statistics were computed using signed areas.

Adjusted *p*-values that are statistically significant (< .05) and NCDIF values that are above or close to the cutoff values are in bold face.
For dichotomous items with WSC scores, cutoff points for NCDIF were calculated using the Monte Carlo item parameter replication procedure (Cervantes, 2012; 2017) with 1000 replications. For polytomous items, a cutoff value of 0.096 was used for items with five response categories, and 0.15 was used for items with six response categories (Oshima & Morris, 2008; Raju et al., 1995).
NCDIF power indicates the post-hoc power to detect the true NCDIF for each item, considering the sample sizes and ability distributions. Power was reported for dichotomous items only, currently supported by the package.

Figure 1. *Literature Review: Literature Search and Selection Process*

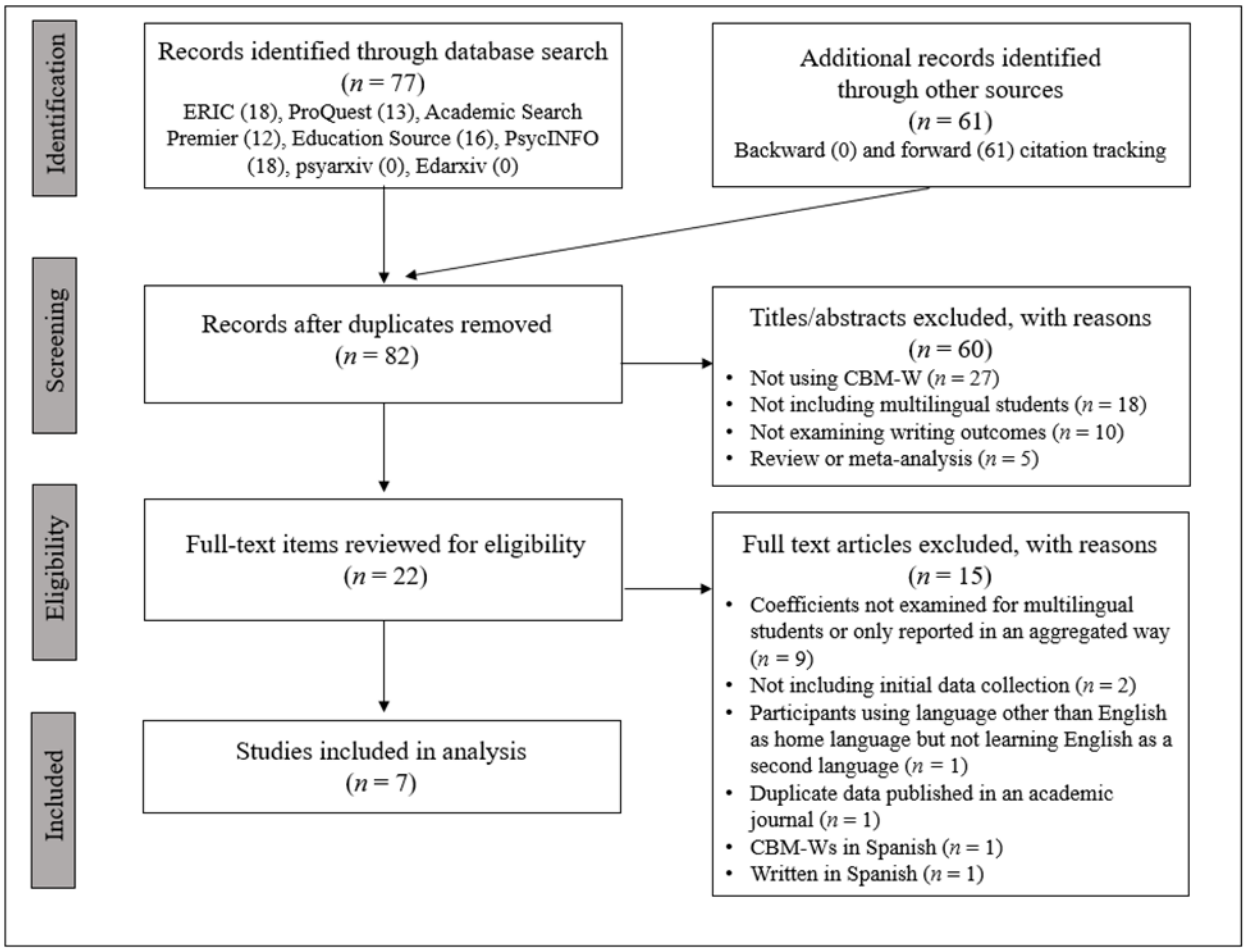
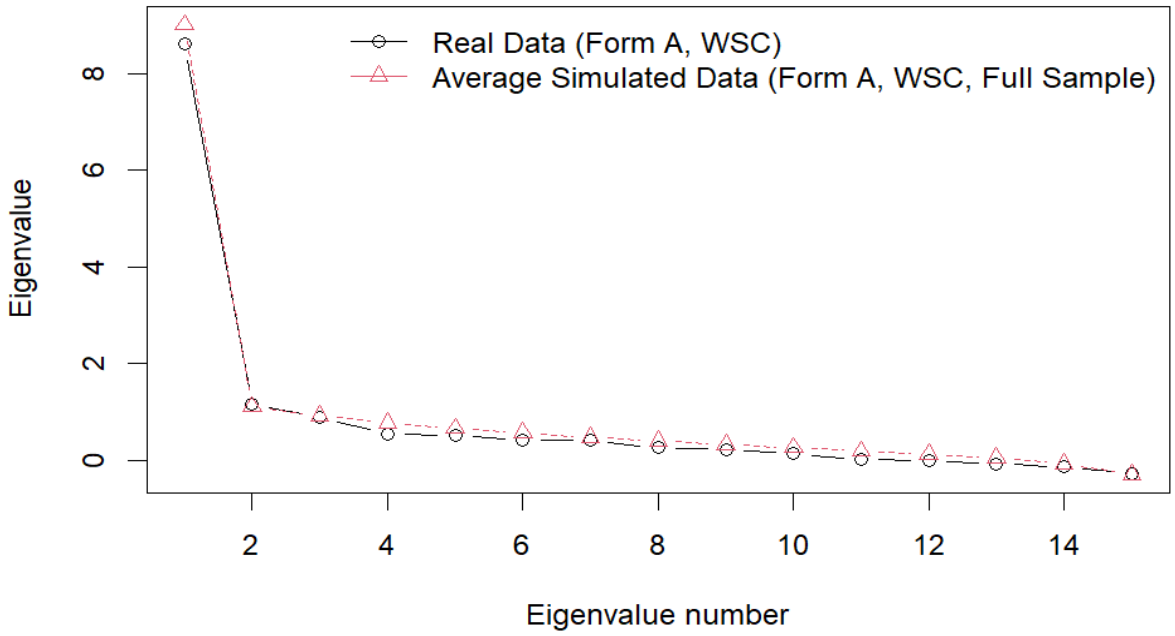


Figure 2. *Scree Plots Showing Eigenvalues of Observed and Simulated Data under the Unidimensional Rasch Model (Words Spelled Correctly)*



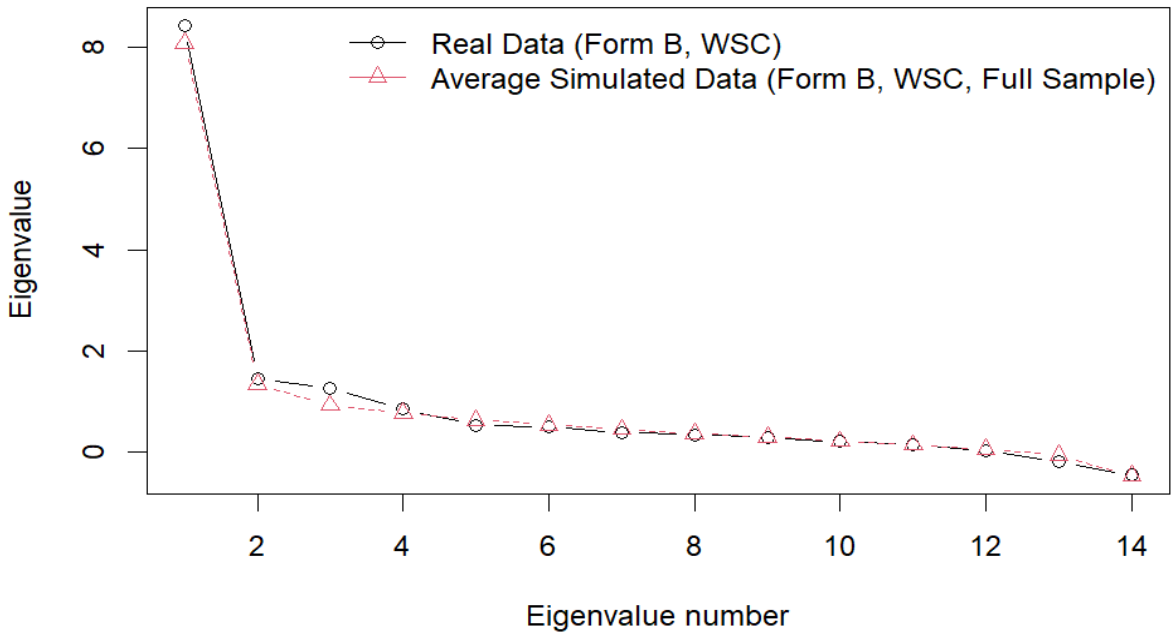


Figure 3. *Scree Plots Showing Eigenvalues from the Exploratory Factor Analysis Using Principal Components Estimation for Polytomous Items (Correct Letter Sequences)*

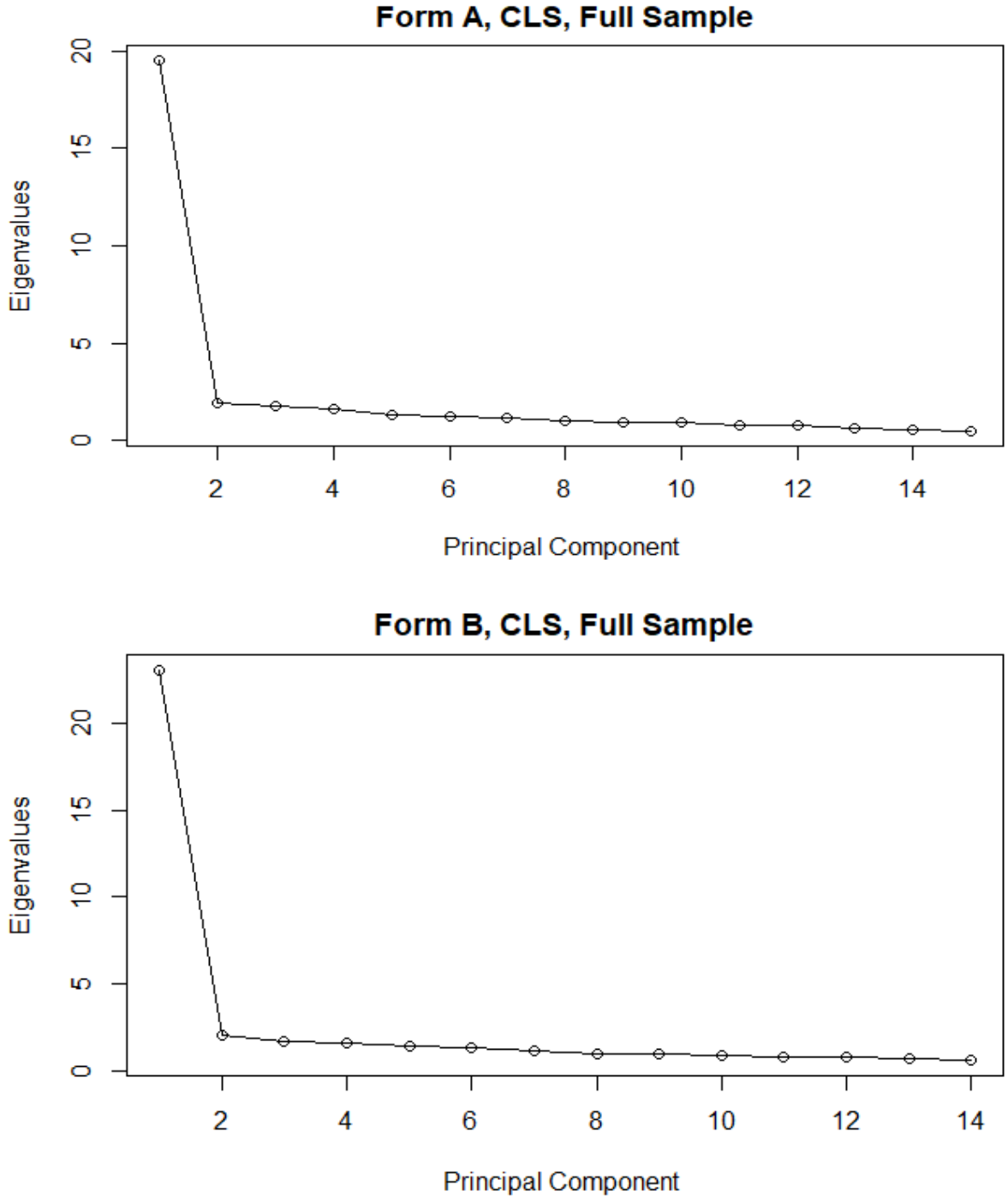


Figure 4. *Item Response Functions for Form A, Using Words Spelled Correctly*

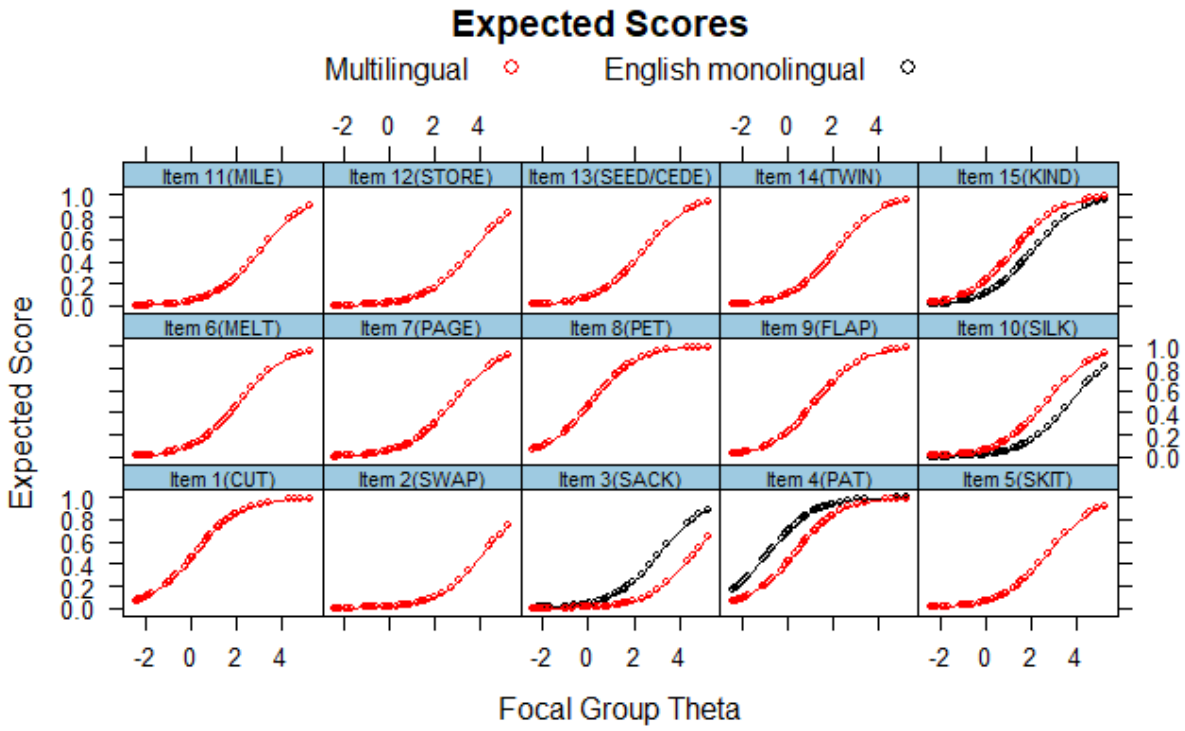


Figure 5. *Item Response Functions for Form B, Using Words Spelled Correctly*

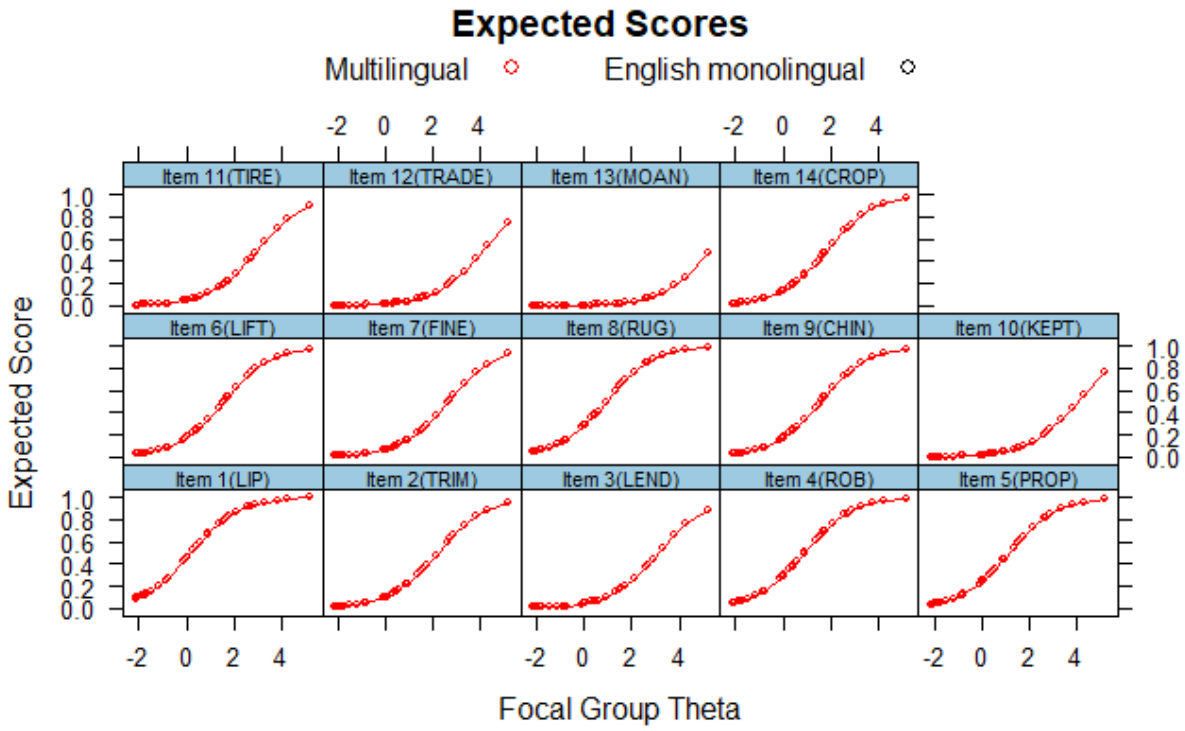
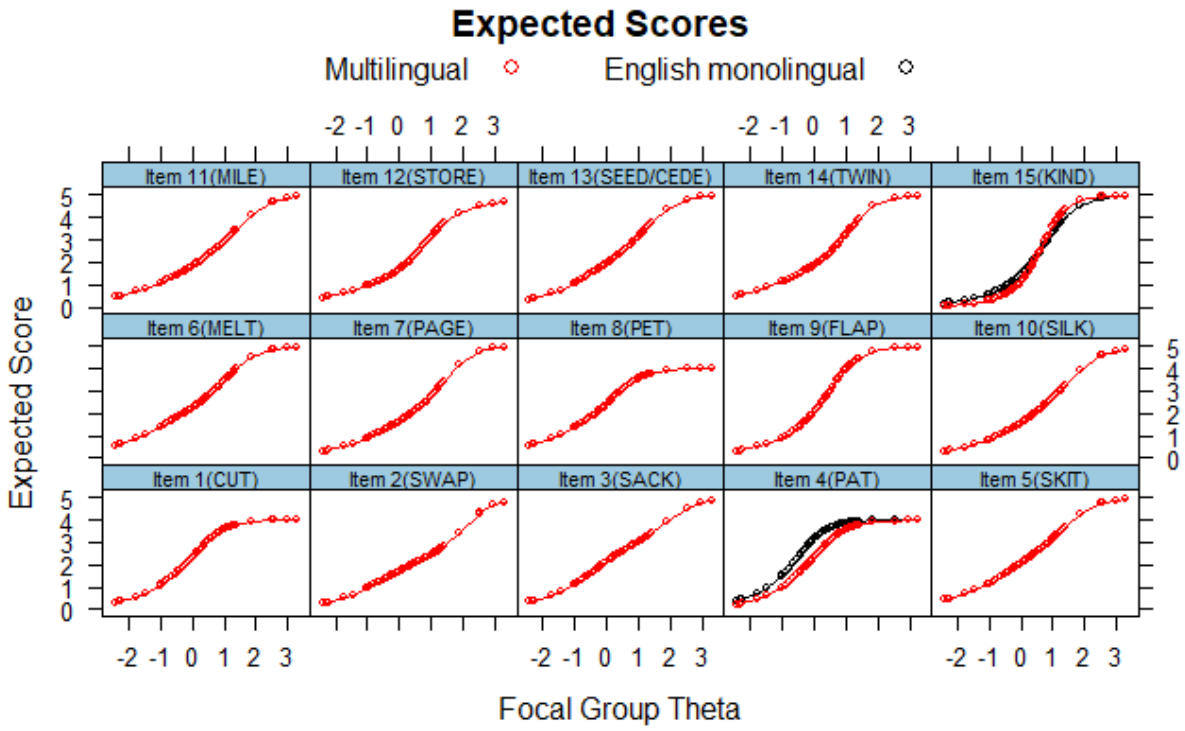
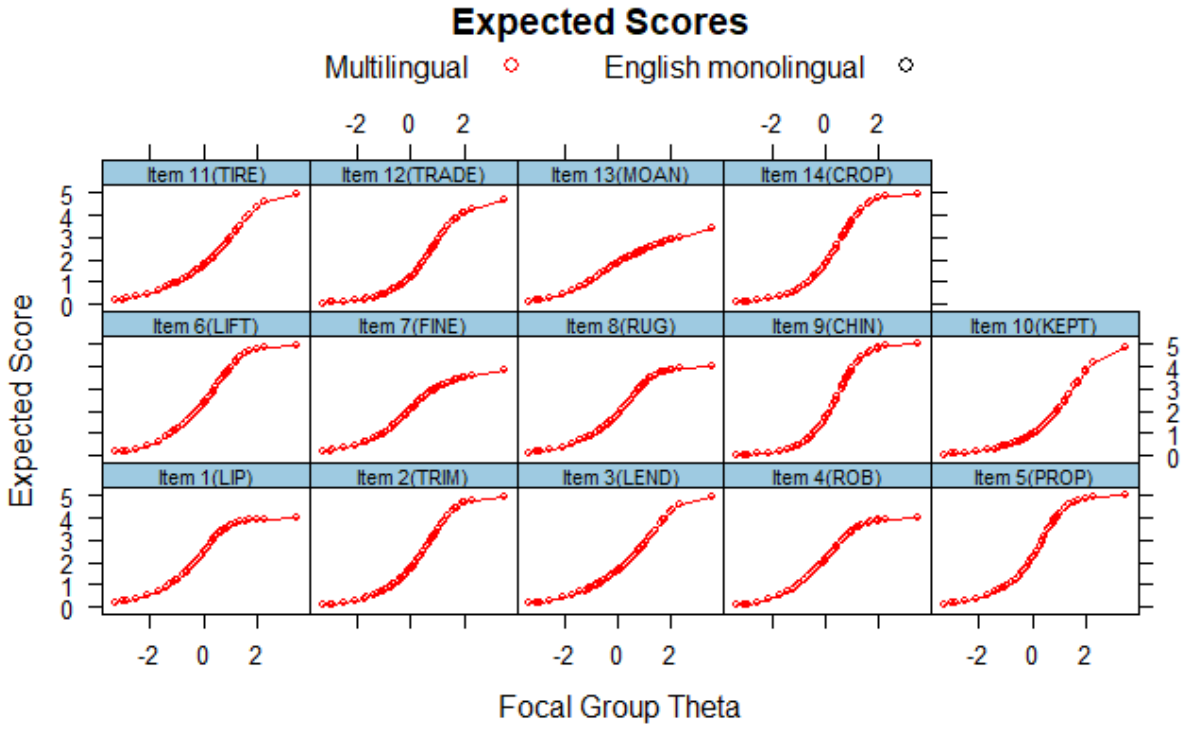


Figure 6. *Item Response Functions for Form A, Using Correct Letter Sequences*



Note. Item 12 ("STORE") was intended to have seven response categories (0 to 6 CLS scores). However, none of the students received a CLS score of 5, resulting in only five response categories being examined.

Figure 7. *Item Response Functions for Form B, Using Correct Letter Sequences*



Note. Item 12 ("TRADE") was intended to have seven response categories (0 to 6 CLS scores). However, none of the students received a CLS score of 5, resulting in only five response categories being examined.

Appendix A
Word Dictation Administration

Reference: McMaster, K. L., & Lembke, E. S. (2018). *Data-based instruction in beginning writing: A manual.* Minneapolis. MN: University of Minnesota.

Materials Needed:

1. Timer
2. Pencils
3. Directions for administration
4. Teacher copy of Word Dictation task
5. Student copy of Word Dictation task

Directions:

This task must be administered individually to the student. Say to the student: ***Today you will write some words for me. I will read each word two times, and then you will write the word on your paper. It's okay if you don't know how to spell a word. Do your best and then we will move on to the next word. Let's start with a practice word. Write the word "cat" on your paper. "Cat."***

Monitor the student to see that he/she is writing the word on the top line of his/her paper under "Example." Don't worry about spelling mistakes. When the student is finished or pauses for more than 5 seconds on the practice word, demonstrate how to write the word on the line.

Now, you will write some more words. When you are finished with one word, move down a line and get ready for the next word. If you make a mistake, just cross it out. Do you have any questions? Remember to do your best! (Set timer for 3 minutes)

*Here is your first word…*_____. Start timer *after* you say the first word. Beginning with the first word, say each word two times, pausing briefly in between. Do not say anything else, such as "The next word is…" or "Number 2 is…" Go on to the next word when the student is finished, or when the student pauses on a word for more than 5 seconds, in which case you would say to the student: **"*Try the next word."***

Do not provide any prompts to the student after the initial word reading. Read words at a consistent pace, without rushing the student. Time the student for 3 minutes. If the student finishes before 3 minutes, record the exact time remaining on the student form. If the student is in the middle of writing a word when the timer rings, make a mark behind the last letter written before the timer rings, and score accordingly.

When the timer rings, say ***Stop. Thank you for working so hard!***

Shortened Directions for progress monitoring:

Say: ***Now we will write some words. I will say each word two times and you will write it. When you are finished writing a word, move down a line and get ready for the next word. Remember to do your best!*** (Set the timer for 3 minutes.) ***Here is your first word...*** Start timer *after* you say the first word. When the timer rings, say: ***Stop. Thank you for working so hard!***

Appendix B
Word Dictation Scoring

Reference: McMaster, K. L., & Lembke, E. S. (2018). *Data-based instruction in beginning writing: A manual.* Minneapolis. MN: University of Minnesota.

Materials Needed:

1. Red and blue colored pencils: Blue = correct & Red = incorrect
2. List of administered words and student packet.
3. Record student name, week, and the date student completed the task.

Scoring Procedures:

For word dictation, count:

1. The number of words written (WW),
2. Words spelled correctly (WSC),
3. Correct letter sequences (CLS),
4. Incorrect letter sequences (ILS).

Words Written (WW)

1. Count the number of words written. A word is defined as a series of two or more letters on a line or separated by spaces on each side.
    a. If the student is in the middle of writing a word when the timer stops, and they have written 2 or more letters, it counts as a word written.
    b. Score only the word that represents your best judgment of what the student meant to write for the target word

Words Spelled Correctly (WSC)

1. A word counts as a WSC ***only*** if it matches the *target word*. If the student spelled another English word but it does not match the target word, it is scored as an Incorrect Word.
   Tip: Score the Word Dictation probes with the list of administered words next to the student response sheet to check answers.
   Example: target word is "drove" but student wrote "drive" **(WSC = 0)**
2. Underline incorrectly spelled words in red.
3. Calculate WSC by subtracting underlined words from WW.
4. Reversals of correct letter formation would cause the word to be scored as incorrect.
   Example: cat**ɐ**. **(WSC = 0)**

Correct Letter Sequences (CLS) and Incorrect Letter Sequences (ILS)

1. A correct letter sequence is one that contains any two adjacent, correctly placed letters.
2. Use the caret method for scoring. Place a blue caret ^ above two letters if it represents a correct letter sequence, and a red caret **v** below the letters if it represents an incorrect sequence. Score incorrect sequences first using a red pencil below the line. Then score correct sequences with a blue pencil above the line.
3. Score letter sequences:
    a. Score a correct letter sequence at the beginning of the word if the first letter of the word is correct.

   b. Score an incorrect letter sequence at the beginning of the word if the first letter is incorrect.
   c. Continue to score correct and incorrect sequences through the rest of the word.
   d. Score a correct sequence at the end of the word if the last letter is correct.
   e. Score an incorrect sequence at the end of the word if the last letter is incorrect.
4. If student is in the middle of writing a word when the timer rings, score the letter sequences written up to the last letter. Do not score the final sequence as either correct or incorrect.
   Example: ^c^l^a
   (WW = 1; WSC = 0; CLS = 3; ILS = 0)
5. If a word ends in a double letter (e.g., grass), and the student writes the word with only one letter, the sequence at the end of the word is scored with one incorrect letter sequence. The word would not count as a word spelled correctly. Consider the following examples (dictated word = grass):
   Example: **v**r^a^s**v**
   (WW = 1, WSC = 0, CLS = 2, ILS = 2)
   Example: ^g^r^a^s**v**
   (WW = 1, WSC = 0, CLS = 4, ILS 1)
6. If a student omits a letter in the middle of a word, score with one incorrect letter sequence. Consider the following example where the student wrote *wed* for *weed.*
   Example: ^w^e**v**d^
   (WW = 1, WSC = 0, CLS = 3, ILS = 1)
7. If a student doubles a letter inside a word, but otherwise has spelled the word correctly (e.g., classp for clasp), score an incorrect letter sequence on either side of the second double letter.
   Example: ^c^l^a^s**v**s**v**p^
   (WW = 1, WSC = 0, CLS = 5, ILS = 2)
8. Count total correct sequences. Count total incorrect sequences.

Prorate When Student Completes Task Before Time Limit

If a student completes all 30 words on this task prior to the 3-minute time limit, you must record the exact time that the student finished writing the final word or letter sequence. Next change the completion time to seconds (i.e., 2 minutes, 40 seconds = 160 seconds).
Use this formula:
Number correct (WW, WSC, CLS) divided by seconds it took the student to complete the task and multiplied by 180 (total time allowed to complete task).

Example: If a student finished the Word Dictation task in 160 seconds and obtained 120 CLS, you would divide 120 by 160 and then multiply by 180. This would adjust the number of CLS to 135.

This same formula works for each scoring technique: WW, WSC, and CLS.